TECNOLOGIAS DE LA INFORMACION Y LA COMUNICACION AVANZADAS PARA EL DESARROLLO SOSTENIBLE Y LA INNOVACION

CATAÏ WORKSHOP
SELECTED CONTRIBUTIONS

CATAI
Collaboration France - Colombia

SELECTED CONTRIBUTIONS

CATAÏ, Collaboration France-Colombie en Technologies de l'Information et la Communication Avancées pour le Développement Durable et l'Innovation - Colaboracion Francia-Colombia en Tecnologias de la Informacion y la Comunicacion Avanzadas para el Desarrollo Sostenible y la Innovacion
www.catai.fr

Multilingual Book: English, French, Spanish
Lyon, France 2025

9791098393600

ISBN 979-10-983936-0-0
EAN 9791098393600

EDITION

- **Oscar CARRILLO, CPE, INSA** Lyon, Inria, CITI, UR3720, 69621 Villeurbanne, France
- **Carlos Jaime BARRIOS HERNANDEZ,** SC3 UIS, Bucaramanga, Colombia, LIG/ INRIA Grenoble and INSA Lyon, Inria, CITI, UR3720, 69621 Villeurbanne, France.
- **Fréderic Le MOÜEL**, INSA Lyon, Inria, CITI, UR3720, 69621 Villeurbanne, France

ACADEMIC COMMITTEE LEADERSHIP

- **José Tiberio HERNANDEZ PEÑALOZA**, Universidad de los Andes, Bogotá, Colombia
- **Yves DENNEULIN**, Grenoble INP, ENSIMAG, LIG/INRIA and Persyval, Grenoble, France
- **Michel RIVEILL**, UCA, INRIA Sophia Antipolis, France
- **Frédéric MERIENNE**, ENSAM-Paris Tech, Chalon-sur-Saône, France
- **Claudia RONCANCIO**, Grenoble INP, ENSIMAG and LIG, Grenoble, France
- **Harold Enrique CASTRO BARRERA**, Universidad de los Andes, Bogotá, Colombia
- **Helga DUARTE**, Universidad Nacional de Colombia, Bogotá, Colombia
- **José Ismael PEÑA**, Universidad Nacional de Colombia, Bogotá, Colombia

SPECIAL THANKS

- Alliance Française de Bucaramanga, Colombia
- Ambassade de France en Colombie, Colombia
- Embajada de Colombia en Francia
- Centre d'Innovation en Telecommunications et Integration des Services - CITI, France
- École Supérieur de Chimie, Physique et Électronique de Lyon, CPE Lyon, France
- Institut National de Sciences Appliquées de Lyon – INSA Lyon, France
- Université de Grenoble Alpes, UGA, France
- Université Côte d'Azur, UCA, France
- École Nationale Supérieur d'Arts et Métiers - ENSAM, Chalon sur Saône, France
- Universidad Industrial de Santander - UIS, Bucaramanga, Colombia
- Parque Tecnológico de Guatigaura, Piedecuesta, Colombia
- Supercomputación y Cálculo Científico, SC3UIS, Colombia
- Universidad de los Andes - UniAndes, Colombia
- Universidad Nacional de Colombia - UNAL, Colombia
- Laboratoire d'Informatique de Grenoble - LIG, France
- Institut National de Recherche en Informatique et Automatique - INRIA, France
- ATOS, France
- Alcaldía de Bucaramanga, Colombia
- Asociación Recreativa de Profesores de la Universidad Industrial de Santander y personal de la sede recreativa Catay en Piedecuesta, Santander, Colombia.
- Sistema de Cómputo Avanzado para América Latina y el Caribe, SCALAC.

DESIGN AND LAYOUT

www.wirebgraphics.com

BOOK PUBLISHED BY

Éditions CPE (Lyon)

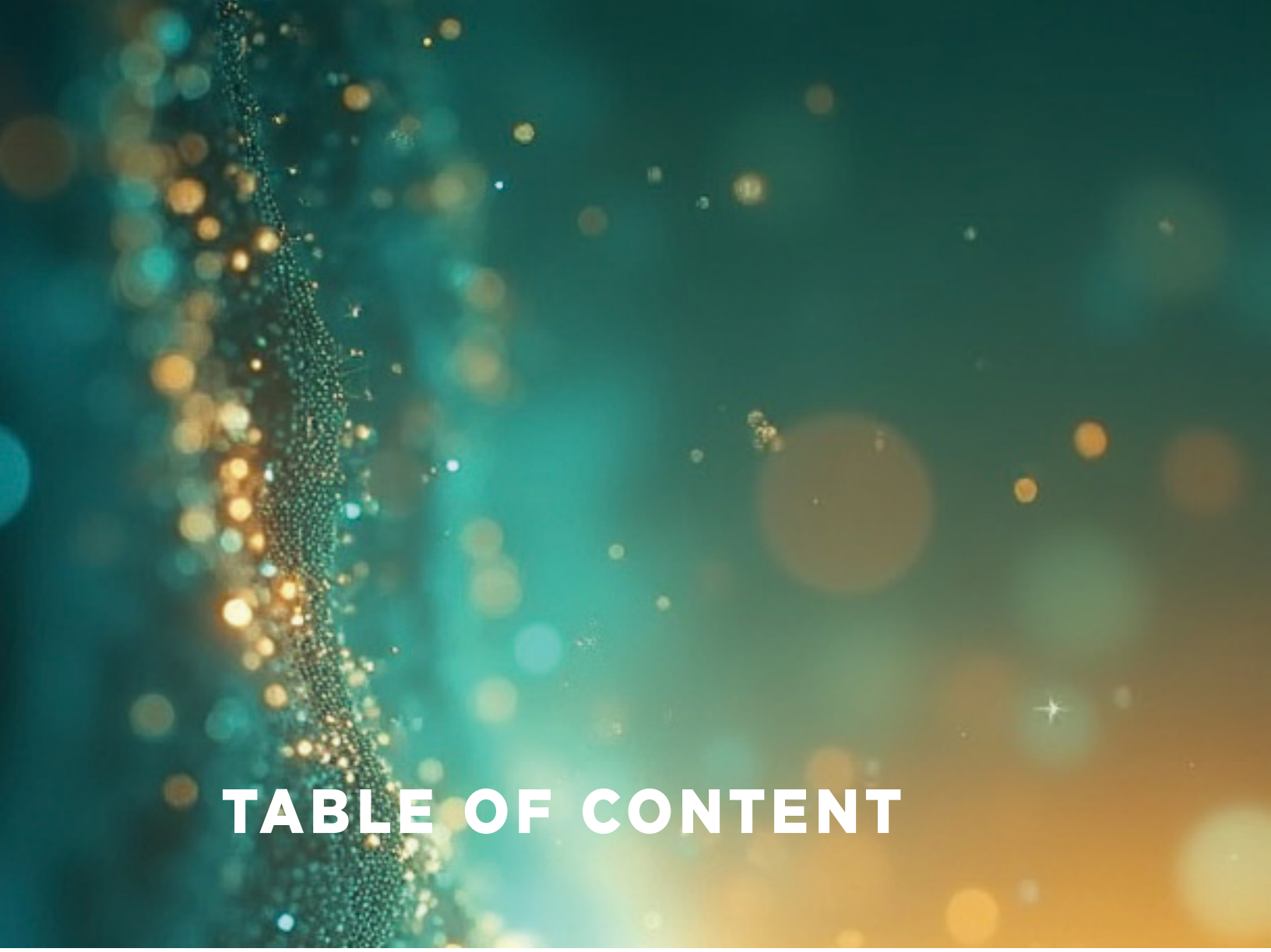


TABLE OF CONTENT

ABSTRACTS

CIUDADES INTELIGENTES PARA LA TOMA DE DECISIONES

GINA PAOLA MAESTRE GÓNGORA

Las ciudades inteligentes no se definen únicamente por el uso de tecnologías como la inteligencia artificial o los sensores, sino por su capacidad para mejorar la calidad de vida de los ciudadanos a través de decisiones más informadas y servicios públicos más eficientes. En Colombia, iniciativas impulsadas por el Ministerio TIC y proyectos desarrollados en ciudades como Bogotá y Medellín muestran avances significativos. Sin embargo, persisten desafíos importantes como la brecha digital, las desigualdades en infraestructura y la necesidad de una regulación más robusta.

A pesar de estos retos, el compromiso de universidades, empresas y ciudadanos ha dado lugar a soluciones innovadoras basadas en la colaboración. Para construir territorios verdaderamente inteligentes e inclusivos, es fundamental que la transformación digital no quede exclusivamente en manos del Estado o del sector tecnológico, sino que integre de forma activa a la ciudadanía. Los beneficios de este enfoque son evidentes: mejor movilidad, mayor seguridad y una mejor calidad de vida.

Uno de los aportes más valiosos de las ciudades inteligentes es su capacidad para tomar decisiones fundamentadas en datos. El uso de big data, inteligencia artificial y plataformas digitales permite comprender en tiempo real las necesidades de las comunidades, optimizar el uso de recursos públicos y anticipar problemas antes de que se agraven. Esto fortalece la gobernanza y permite formular políticas públicas más eficaces, centradas en las personas.

Colombia tiene el potencial de convertirse en un referente regional si adopta un modelo participativo, sostenible y basado en la inteligencia colectiva y en la toma de decisiones basada en evidencia

GINA PAOLA Maestre Góngora
Profesora Universidad de Antioquia
Departamento Ingeniería de Sistemas, Facultad de Ingeniería
Oficina 21-428, Medellín
E-mail: gina.maestre@udea.edu.co

MISSION PLANNING WITH MULTICONSTELLATION SATELLITES FOR CRISIS MANAGEMENT

MICHAEL S. PUENTES

CARLOS J. BARRIOS

OSCAR CARRILLO

UNIVERSITÉ DE LYON, INSA DE LYON, LYON, FRANCE

ABSTRACT

Positioning using a GNSS (Global Navigation Satellite System) service is only possible with the simultaneous reception of the signal from at least four satellites.

The probability of loss of visibility of GNSS satellites in urban environments is especially critical due to the architectural elements typical of a city, making it impossible to locate a receiver. This will be crucial in situations of alert for collisions or that requires continuous monitoring of the position.

This paper describes a method of planning urban routes according to the needs of positioning associated with decentralized systems for emergency management in urban environments, integrating the positions calculation of satellites through orbital mechanics, with images of the buildings in the urban environment and semantic segmentation with Deep Learning techniques.

In this way, predict the visibility relation between an observer and the satellites of a GNSS service. The satellite unavailability due to the loss of sent signal can be replaced by decentralized systems within the city, which are exempt from indisposition by an elevated architecture, but it can present nearby obstructions and limited signal reach, for which it is necessary to have a greater source of local services (crowdsensing/crowdsourcing). In synergy with the ALERT project supported by CITI laboratory from INSA Lyon and SC3 from Universidad Industrial de Santander, for an implementation in Bucaramanga's city.

Keywords: Smart Cities, GNSS, Decentralized services

Michael S. Puentes | Carlos J. Barrios
Universidad Industrial de Santander, Bucaramanga, Colombia
Oscar Carrillo
Université de Lyon, INSA de Lyon, Lyon, France

SHORT PAPERS

COST EVALUATION OF TRANSPORT ANALYTIC IN PUBLIC CLOUDS

PEDRO PREZ[1], CRISTIAN CASTELLANOS[1], YVES DENNEULIN[2] AND HAROLD CASTRO[1]

[1]*SYSTEMS AND COMPUTING ENGINEERING DEPARTMENT, UNIVERSIDAD DE LOS ANDES, CRA 1ESTE 19A 40, BOGOTA, COLOMBIA* [2]*ENSIMAG, INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE, GRENOBLE, FRANCE*

{PPEREZ, CC.CASTELLANOS87, HCASTRO}@UNIANDES.EDU.CO, YVES.DENNEULIN@GRENOBLE-INP.FR

Keywords: Lambda architecture, cost comparison, performance evaluation, transport analytics, bus delay prediction, Software as a Service

Abstract: The world is experimenting with a data deluge every industry is struggling to embrace. Processing both stream and batch data is a new requirement for several use cases in analytics. In this paper, we present a cost comparison of using three different providers to process transport data. We evaluate Google Cloud Platform, Microsoft Azure, and Amazon Web Services Cloud. The evaluation is carried out comparing performance and costs in a public transportation delay monitoring case study assessing various concurrency scenarios.

1. INTRODUCTION

The public cloud is a natural environment to implement big data analytics solutions. Lambda architecture (Marz and Warren, 2015) represents a well known pattern for real-time and batch data processing. Applying such technologies to particular industries and measure the impact of utilizing the different services available on leader Cloud Service Providers (CSP) will boost the adoption of data analytics on those industries.

Delay monitoring in public transportation is a good example of a use case requiring such analysis. Its services require combining the processing of large datasets of vehicle location and low latency to report the delay times to users in near real-time. The paper is organized as follows: in section 2 we introduce the case study of transportation analytics. Section 3 describes the different implementations of Lambda architecture using SaaS. Section 4 reports and discusses the results obtained. Finally, section 5 outlines the conclusions.

CASE STUDY

A description of the main services provided in an Intelligent Transport System (ITS) are specified in the ISO 14813-1 standard (ISO, 2001). There are five main methods to determine bus arrival time predicral Networks (ANN) models (Kumar et al., 2014; van Hinsbergen et al., 2009; Chien et al., 2002). The other four main methods are: The Kalman filter (Shalaby and Farhan, 2004; Chen et al., 2004), support vector machine

(Hernandez, 2014; He et al., 2013; Yu et al., 2011; Bin et al., 2006), regression analysis models (Jeong and Rilett, 2005) and time series models (Sapankevych and Sankar, 2009).

Our case study presents a proposed bus arrival time prediction with Lambda architecture. The developed architecture covers the batch layer using historical data with one day execution window, and the speed layer uses real-time data with five minutes execution window continually during the day. The algorithm in both layers is an expected average delay in five-minute windows. These windows are generated for each key composed of route id, stop id and window time. Additionally, delay average is grouped by day of the week. The window time is defined by the groups of trip updates reported within five minutes.

We take the Metro Vancouver's regional transportation (Translink) GTFS data set which is publicly available (Translink Open API [1]) and real-time Trip Update data (GTFS real-time Open API) which provides real-time Vancouver's transportation data for the analysis. [1]https://developer.translink.ca/

2.1 TRANSLINK DATA SET

The open API of Translink serves trip updates data in GTFS realtime (protobuffer format), and we send requests to collect feeds every sixty seconds. These data are collected during one week from December 11, 2017 to December 17, 2017, and 16 hours every day. The GTFS real-time data contains just over 6,720 trip updates with 4,631,075 protobuffer files which are deserialized to JSON format.

In summary, these JSON files comprises 211 routes and 8,447 stops, each pair with a delay time to the next stop. The size of the data set (binary format) is 383 MB in 6,720 individual files. Each trip update contains the information presented in the appendix section at the end of the article.

2.1 STEPS NEEDED TO CALCULATE THE WAITING TIME

A Trip Update provides information in real-time about the trips in operation in the city of Vancouver. This means that the first step is to join the planned trips file in GTFS file with each Trip Update in GTFS real-time. This step is necessary in both layers.

In the next step, the speed layer receives every sixty seconds a Travel Update with approximately 45,000 JSON updates. The algorithm makes groups every five minutes (time window) with exactly five JSON updates. Then the speed layer assembles tuples with route id, stop id, and its expected delay average.

At the end, the speed layer write each five minutes the results composed by stop id, route id, week day, time window and avg delay in the serving layer. Consequently, the preprocessed view with real-time information calculation is ready to respond users requests. The goal is to guarantee new data available as soon as needed for the user queries thus offering real-time views.

Simultaneously, the batch layer job is executed at the end of the day to compute whole stored raw data generating the same output (stop id, route id, week day, time window and avg delay). The batch layer writes each day the results over serving layer recomputing cumulatively historic data. This heavy workload implies high latency processing, and therefore the speed layer compensates this limitation.
Lastly, we implement and evaluate the lambda ar Implementation

3 IMPLEMENTATION

To compare each lambda architecture SaaS instance, we implement versions for each cloud platform regarding the stack of SaaS offered by each public cloud vendor (Amazon, Google and Azure). In each layer of Lambda architecture, we select the service with the highest level of abstraction and the best Service Level Agreements (SLA) in terms of availability and performance. This selection is made for two main reasons: to avoid low level implementation and to make the metrics comparable.

3.1 AWS IMPLEMENTATION

The AWS object storage provides security, compliance capabilities, flexibility and cost optimization. Speed layer is implemented in Amazon Kinesis Data Analytics because it facilitates the streaming processing in real time with standard SQL. Regarding the batch layer, AWS Glue is a fully managed ETL[2] service that simplifies the preparation, manipulation and loading of data through scheduled jobs. The views reside in AWS Redshift which is a fast, fully managed and columnar storage data warehouse which eases and cheapens the data querying using standard SQL parallel execution.

3.2 GOOGLE CLOUD IMPLEMENTATION

In Google cloud the GTFS-realtime ingestion is implemented using Cloud Pub/Sub for exploiting the scalability, flexibility, reliability and low-latency service which allows to consume messages asynchronously from external data sources. Cloud Data flow service combines both the batch and streaming pipelines in a unified programming model, therefore the batch and speed layers fit the characteristics of this service. The master dataset is realized with Object Storage service which is designed for secure, durable, cheaper and raw storage.The outputs of cloud Data flow, batch and speed views, are materialized in Google Cloud Datastore, a NoSQL document database which offers automatic scaling, high performance, and SQL-like query language.

3.3 AZURE IMPLEMENTATION

For Azure implementation, we use the Event Hubs service that allows to support distributed streaming and batch ingestion of GTFS-realtime binary files using a publish-subscribe schema. Azure Data Lake architecture using SaaS with a realistic and exhaustive tests described in the next Sections.[2]Extract, Transform, Load

Table 1: Summary of data processed by the Batch layer

Days	Trip Updates	Processed JSON	Accumulated percentage
1	1,177	722,958	15.6%
2	2,161	1,490,966	32.2%
3	3,114	2,200,500	47.5%
4	4,150	2,914,376	62.9%
5	5,064	3,650,607	78.8%
6	6,048	4,182,829	90.3%
7	6,720	4,631,075	100.0%

Store is the service recommended for big data analytics workloads. It is secured, HDFS compliant which allows to run massively parallel analytics, for this reason it is ideal as Master dataset. Batch layer is covered by Data Lake Analytics, its highest level service performs heavy workload jobs, writing queries which scale instantly. On the other hand, speed layer is performed on Stream Analytics service to enable run massively parallel real-time analytics data streaming using SQL–like language. The views repository created in both speed and batch layers is implemented in Cosmos DB, a database for low latency and scalable querying with support for NoSQL.

4 EVALUATION

For this case study, we focus on evaluate the performance and cost of each public cloud. The results of this evaluation are described in this section.

4.1 PERFORMANCE

The performance test of the batch layer involves the cumulative processing of trip update files each day, these file numbers are shown in Table. 1. We can also determine the number of JSON that were processed every day by the batch layer.

Before starting the processing in the batch layer, the raw of Trip Updates should be read, for this reason Figure 1 presents the average reading time for each implementation. The average reading time of AWS Glue in AWS S3 storage is the most stable and efficient, while the other batch services take 12 times (Google Cloud) and 18 times (Azure) more time reading raw data.

The average reading time of the Cloud Datastore service in Google Cloud has a constant increase as the number of Trips Update increases every day. And finally, the average reading time of the Data Lake Store service in Azure has the highest increase until the fifth day, after that day the average reading time has a decrease, which may reflect a scaling of the service.

After reading the files the next step is to calculate the waiting time. This processing time is shown in

Days	Number of Trip update files	Google Cloud	AWS	AZURE
1	1,177	15	4.1	33
2	2,161	26	4.2	63
3	3,114	41	4.2	86
4	4,150	55	4.4	102
5	5,064	60	4.1	133
6	6,048	70	4.3	79
7	6,720	95	4.3	57
Total	28,434	362	30	553

Figure 1: Average reading time in seconds to Batch layer

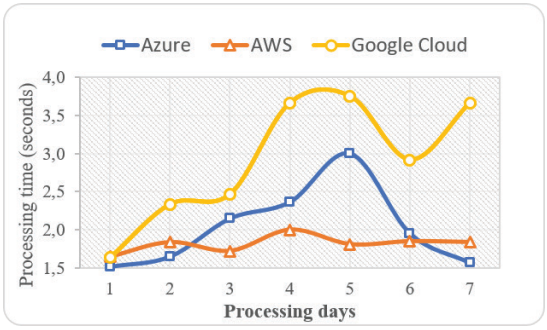


Figure 2: Average processing time for Batch layer

Figure 2. The AWS Glue service that does the processing of the batch layer in AWS, again is the most consistent and efficient, since the processing time is almost constant below two seconds in each execution, despite the increasing amount of files. In contrast, the Google Cloud Dataflow service has the lowest processing performance with peaks almost four seconds, twice the processing time of AWS. Data Lake Analytics in Azure is the most sensitive to the number of processed files, and similar to reading time the service seems to have scaled during the fifth and sixth day.

The final step of batch processing is to write in the serving layer. The average writing time is shown in Figure 3. Amazon S3 service continue with consistent behavior offering the best performance. Conversely, Google BigQuery presents the worst average writing times describing a decreasing trend. In addition, Cosmos DB service presents intermediate average writing times with a slight degradation observed in the last two days. The processing times obtained in speed layer are constant in all platforms constrained to real-time windows, and for this reason we do not consider valuable to compare them.

Days	Number of records	Google Cloud	AWS	AZURE
1	571,917	88	54	62
2	1,024,509	102	39	91
3	1,460,840	145	47	111
4	1,934,729	196	49	68
5	2,303,636	245	46	118
6	2,642,347	326	40	232
7	2,860,524	380	42	239
Total	12,798,502	1482	318	921

Figure 3: Average writing time in Batch layer

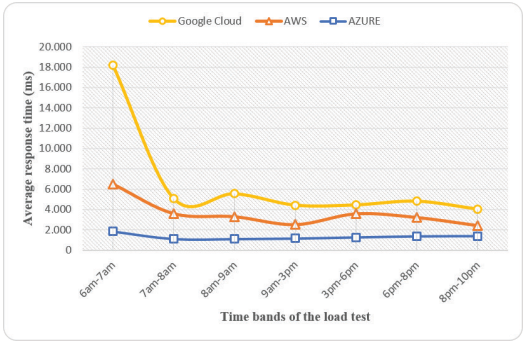


Figure 4: Average response time for Serving layer

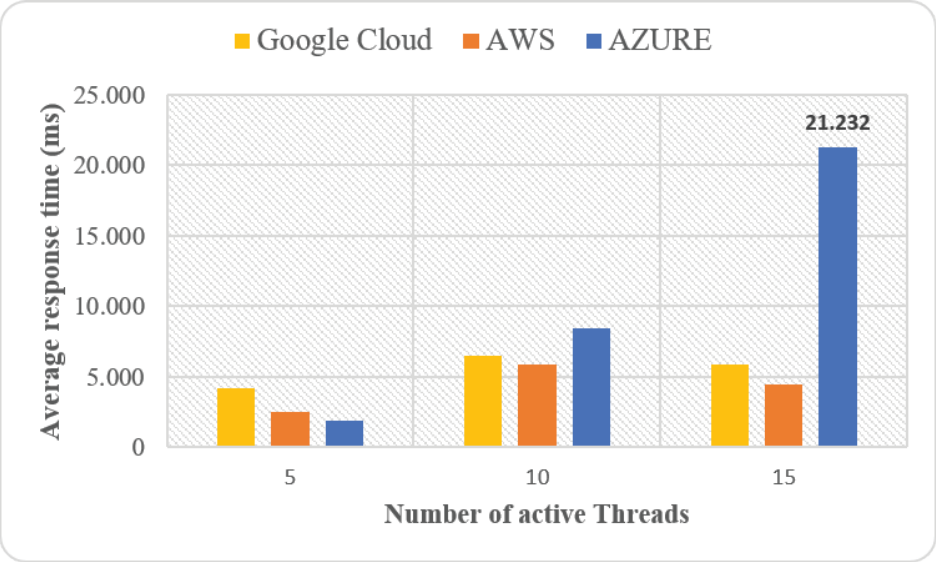


Figure 5: Average response time vs number of active threads

The metric of serving layer performance in respect of response time is shown in Figure 4. It is worthy of note that at beginning of the stress test, all services start with the highest latency specially noticeable in Google serving layer, but when test moves forward, the latency is reduced. Cosmos DB depicts the lowest average response times, followed by AWS Athena and Google BigQuery respectively.

Despite the fact that Azure Cosmos DB service produces the lowest average response time, Figure 5 shows that it rapidly degrades its response time when request's concurrency increases. On the contrary, Amazon S3 and Google BigQuery support concurrency growing without affecting their performance significantly.

To obtain the average response time of 21 seconds generated by Cosmos DB with 15 active threads, Amazon S3 would need 56 concurrent threads, and Google BigQuery would require 40 active threads. Therefore, Amazon S3 is the Serving layer that best supports concurrency without worsening its average response time.Figure 6: Infrastructure monthly costs (USD)

4.2 SERVICES COST

Each implementation of the Lambda architecture is deployed in different public cloud providers, we define and calculate the costs required to replicate a similar case study, with data similar to Vancouver's transportation system and operate them for 4 weeks. As a result, Figure 6 presents a summary of the monthly fees generated by each provider during the simulation.

The highest monthly cost is generated by Azure, and specifically due to the high cost of the Cosmos DB service. AWS Glue and Kinesis, both in batch and speed layer respectively, are the highest individual costs in these layers with respect to the other infrastructures.

Google cloud is the least expensive provider in all layers and with remarkable difference. Finally, regarding the learning curve, Google Cloud free tier enables an inexpensive proof of concept with these SaaS compared to other vendors free tier.

CONCLUSIONS

This document presents a comparison of costs in development, and deployment the same case study over Lambda architecture using three different public cloud providers (Google Cloud, Microsoft Azure, and Amazon Web Services) with the main goal of identifying how different public cloud providers with the same architecture deployment can affect the infrastructure cost of running and performance with concurrence users. So as to get valid results, we implementing three version of the Lambda architecture and deploy each one using a different public cloud provider.

As a result of the developing and testing process of the three implementation deployed, we could understand the challenges that must be overcome to use the Lambda architecture.

In terms of performance, AWS obtained the best metrics in batch and speed layer. In batch layer, AWS reported the best performance in reading, processing and writing time, whereas Google Cloud seems to be affected by increasing data size. Focusing on serving layer performance, Azure presented a constant and efficient behavior, but it shown degradation of response time under heavier load compared to other competitors.

In terms of cost services, Azure was the most expensive provider in serving layer, whereas AWS consumed more credits in Serving Layer due to Cosmos DB service. In contrast, Google Cloud presented the lowest price in all layers and it offers the best free tier to initiate the training.

In summary, when performance is a strong concern, despite the high cost, AWS (in batch and speed layer) is the best choice and Azure (in serving layer) should be selected to obtain the best response times. Nonetheless, if service pricing is an important constraint, Google Cloud offers the best choice with a factor of 1/4.

REFERENCES

Bin, Y., Zhongzhen, Y., and Baozhen, Y. (2006). Bus Arrival Time Prediction Using Support Vector Machines. *Journal of Intelligent Transportation Systems*, 10(4):151–158.

Chen, M., Liu, X., Xia, J., and Chien, S. I. (2004). A Dynamic Bus-Arrival Time Prediction Model Based on APC Data. *Computer-Aided Civil and Infrastructure Engineering*, 19(5):364–376.

Chien, S. I.-J., Ding, Y., and Wei, C. (2002). Dynamic Bus Arrival Time Prediction with Artificial Neural Networks. *Journal of Transportation Engineering*, 128(5):429–438.

He, Z., Yu, H., Du, Y., and Wang, J. (2013). SVM based multi-index evaluation for bus arrival time prediction. In *International Conference on ICT Convergence*, pages 86–90.

Hernandez, T. (2014). Flex Scheduling for Bus Arrival Time Prediction. *Transportation Research Record: Journal of the Transportation Research Board*, 2418:110–115.
ISO (2001). Intelligent transport systems Reference model architecture(s) for de ITS sector. Part 1: ITS service domains, service groups and services.

Jeong, R. and Rilett, L. (2005). Prediction Model of Bus Arrival Time for Real-Time Applications. *Transportation Research Record: Journal of the Transportation Research Board*, 1927:195–204.

Kumar, V., Kumar, B., and Vanajakshi, L. (2014). Comparison of Model Based and Machine Learning Approaches for Bus Arrival Time Prediction. *Transortation Research Board Conference*, pages 14–2518.

Marz, N. and Warren, J. (2015). *Big Data, Principles and best practices of scalable real-time data systems*. Manning Publications Co.

Sapankevych, N. and Sankar, R. (2009). Time Series Prediction Using Support Vector Machines: A Survey. *IEEE Computational Intelligence Magazine*, 4(2):24–38.

Shalaby, A. and Farhan, A. (2004). Prediction Model of Bus Arrival and Departure Times Using AVL and APC Data. *Journal of Public Transportation*, 7(1):41–61.

van Hinsbergen, C., van Lint, J., and van Zuylen, H. (2009). Bayesian committee of neural networks to predict travel times with confidence intervals. *Transportation Research Part C: Emerging Technologies*, 17(5):498– 509.

Yu, B., Lam, W. H., and Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*, 19(6):1157–1170.

SMART ADAPTATION OF BIM FOR VIRTUAL REALITY, DEPENDING ON BUILDING PROJECT ACTORS' NEEDS

PIERRE RAIMBAUD[1] AND RUDING LOU[1], FLORENCE DANGLADE[1], FRÉDÉRIC MERIENNE[1], JOSÉ TIBERIO HERNÁNDEZ[2], PABLO FIGUEROA[2]

[1] LISPEN, ARTS ET MÉTIERS, INSTITUT IMAGE, CHALON/SAÔNE

[2] SYSTEMS AND COMPUTING ENGINEERING, IMAGINE GROUP, UNIVERSIDAD DE LOS ANDES, BOGOTA,

D.C., COLOMBIA ***PIERRE.RAIMBAUD@ENSAM.EU RUDING.LOU@ENSAM.EU FLORENCE.DANGLADE@ENSAM.EU FREDERIC.MERIENNE@ENSAM.EU JHERNAND@UNIANDES.EDU.CO PFIGUERO@UNIANDES.EDU.CO***

ABSTRACT.

Nowadays, virtual reality (VR) is widely used in the AEC (architecture, engineering and construction) industry. One crucial issue is how to reuse Building Information Modeling (BIM) models in VR applications. This paper presents an approach for a smart adaptation of BIM models for using in VR scene, which follows the needs expressed by building projects actors and where the processes are more automatized than in the traditional way. Therefore first the users' needs must be formalized and then the main adaptation consists in filtering BIM data to keep the necessary ones according to the users' objectives. This approach is applied to a study case of a nursery.

Keywords: BIM, virtual reality, users' needs, model adaptation, automated processing

1. INTRODUCTION

The aim of this research is to provide a new methodology for the usage of Building Information Modeling (BIM) in virtual reality (VR) system, based on the needs of the building project actors (professionals and end users) and which contains more automation in their processes. That is why we could call it a smart adaptation: an adaptation in harmony with the expressed needs and made in the more possible automated way. For this purpose, we focus on preparing adapted VR models by filtering the required data, which will be appropriate for the users.

As a reminder, BIM can be defined as all the methods and processes used during the building life cycle (design, construction and usage), joining all the information in one 3D model, providing great possibilities of collaboration among all the participants of AEC (architecture, engineering and construction) projects. Also, according to Saleh et al. [4]'s study, user evaluations have shown that 3D computer visual materials are better than traditional visual materials to support user's participation in architectural design process. That is why BIM and VR are combined for innovations in architecture.

2. RELATED WORK

In the scientific literature, a crucial BIM advantage is its capacity to provide collaboration between all the project actors, facilitating the share of information. However, one limitation is its static aspect, as Heidari et al. explained [2], facing to a hypothetic smart-BIM and dynamic with free interactions. They propose a prototype between an ordinary BIM and a smart one, which provides a virtual environment with predefined interactions on some scene objects: here is the “game aspect” that VR interactions en- tail. Yan et al. [6] propose a BIM-Game prototype, for interoperability between game engines and BIM. It consists on transferring geometric and non-geometric data in both ways (BIM to VR, VR to BIM) through a crossover module. They implicate the users in the design using the VR interactions, applying their Design-Play process, where Design is a phase for the building definition in BIM and Play another to make decisions collaboratively.

For implicating the end users in the first design phases, Bullinger et al. [1] propose to combine participatory design, where the user takes decisions which help with prototyping, and user centered design, where the user is evaluated when testing immersive prototypes, to improve them. Prerequisites are defined with the users thanks to geometric volume layout simulations. Then detailed models are created (based on simulations and data from the first phase, not BIM) and presented through a projection based 3D stereoscopic large-scale displays for a review between end users and professionals.

Wu et al. [5] choose another approach: a BIM-based educational gaming prototype where user profiles are stored: each user participates and data are collected to improve the game. To conclude, Heydariana et al. [3] use a similar approach using less BIM information and more data simulation, studying user choices and experimenting it through a realistic game engine that allows to make decisions on lights.

To resume, all provide methodologies that give interactions to end users only, where VR models are adapted from BIM for only one specific kind of actors. The users’ needs are mostly obtained through interviews and oral exchanges and are not formalized. Moreover, some processes could be automated in order to simplify the work of the computer graphics designer and assure more fidelity from BIM model. In our research, we propose a method and tools to adapt BIM model to VR, which take into account the diversity of actors (professional, end users) and where the processes are more automatized.

3. METHODOLOGY

As a reminder, here we can observe the traditional current method for adapting BIM to VR scenes (fig 1.).

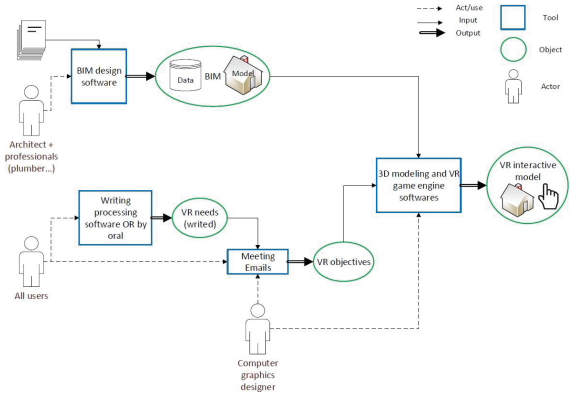


Fig. 1. Traditional methodology schema.

As explained and identified above, this method suffers for some lacks or issues, there- fore we propose the following approach (fig. 2) to intent to solve them.

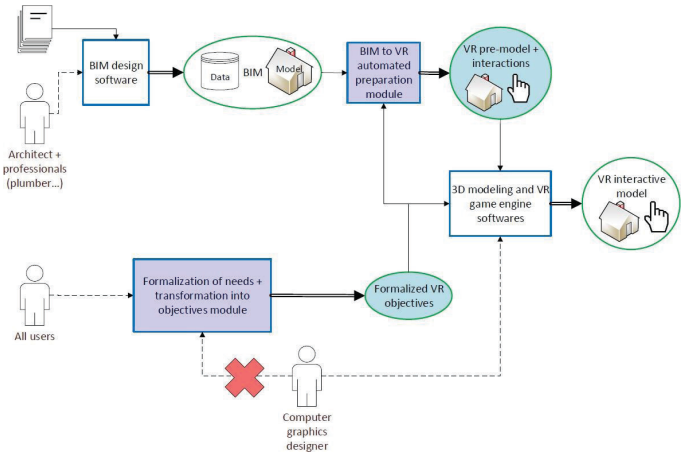


Fig. 2. New methodology schema.

Notice that two new modules have been added, with two new results as output. The first one is a module of formalization of the users' needs and of transformation of these needs in to VR objectives, which will be our criteria to guide the BIM adaptation for VR. The second one is a module of BIM automated preparation to VR, providing then a pre-model of VR with some drafts or models of some interactions.

Therefore, first, we will have to know the users of the VR system(s), their job, their role and the VR usage purpose. Thanks to these information, we could formalized the users' needs and then transform them in criteria. Using them, we could filter the data we want to use from BIM and create a focus on the VR model on what we want to show and what interactions we want to provide thanks to the VR (in terms of content).

For example, in the following case: a building transformation, in built phase, requiring project reviews between the architect and the end users to check the built advancement, would lead us to: end users are the targeted people and the focus extracting information from BIM is on all relevant elements for the building realism.

The emphasis would be on the maneuverability of the virtual walkthrough, about lights and materials.

These criteria would guide the adaptation, to answer to the following needs: the major one is to give a realistic immersion that allows to provide communication about the project between the actors and other ones could be to show the actual state of the construction work and to supply interactions on editable elements.

Then users have to try the model and evaluations are required. Our goal is to ask them how they feel using it and to compare it to a BIM or a not adapted VR model (with all BIM elements).

We propose to use questionnaires; for the previous case, questions could be: how do you rate the maneuverability of the model (navigating, interacting)? And comparing it to a BIM or a not adapted VR model? Was it immersive? Did you feel to be present in the environment? And comparing to BIM? Notice that for this last point we can add an objective measure about the application fluidity, which influence on the immersion quality: the frame rate.

CASE STUDY

This study case is a nursery construction project, a transformation of an old covered market. A division of the council dedicated to the urbanism is responsible of planning and reviewing this kind of projects so we decide to meet the architect in charge.

First, this meeting with the architect allows us to know more about their needs in VR. For this step, we use the traditional method (interview and exchanges by email) in order to define VR needs and then VR objectives. We use this traditional way to identify its lacks and to make choices about our future module for formalization and transformation of the users' needs.

Then, the architect gives us the BIM of the nursery, an Autodesk Revit 2017 ® file. For this step, for this case we mostly apply the traditional method, in the same way than for the first module, in order to identify lacks and issues. First, we have to export it as an FBX file and to open it in Autodesk 3ds Max ®. However, the FBX file is composed of a large amount of elements with no hierarchy. Therefore, we order and rename them to prepare the VR model separating interior and exterior objects.

We also delete the ones that are never visible in a visit. For this case, we just decide to delete the elements under the roof. In 3ds Max, we use ProOptimizer, an integrated tool, which allows to reduce the number of polygons for one object until 70%, due to the high quantity of polygons for this model (2 billion). Notice that materials have been lost when converting the file. We choose for this case to recreate the materials next to the FBX file, but future research could include reflection about BIM materials transfer.

Next, the FBX file is imported to Unity ® to build the VR model with the user interactions. About the metaphor of virtual walkthrough, we choose for this case a teleportation between defined checkpoints, with the render of moving when going to a new one. It allows to navigate easily, in particular outside; however for the indoor visit, other options may be tried in future research because indoor environments often offer more navigation ways so user choices may be more eclectic (check-points could bet not sufficient). Moreover, to overcome the doors issue (open, close) the walk- through has been separated between an indoor and an outdoor visit. The user has to choose before starting, in a helicopter view: it may help, giving an overview, similar to a 2D map but in VR with the visible checkpoints (he can quit the roof).

After that, work has been done for the scene realism (fig. 2) (which include materials work already done), without affecting the fluidity of the VR model, with main ef- forts on the lights and the visibility of the objects. In this study case, we apply a binary choice: when walking outdoor, indoor objects are hidden, to consume less resources and when walking indoor, everything is loaded: in one future work we could show only the rooms close to the current one.

About lights, as the building project is on built phase, major modifications are not possible. But at least users want to check the exposure depending on the sun, the interior lights and the walls/windows materials. For this issue we had to get back the building orientation from Revit to Unity. Then we use the Revit solar cycle simulation to get the positions of the sun path, depending on the date of the year.

In Unity, we simulate three different solar cycles (equinox, summer and winter solstice). So we allow end users to check the building exposure in the VR system and to ensure quality to the user experience, users can stop at any moment the solar cycle and stay with the current exposure.

Fig. 3. Nursery in VR.

About interior lights, windows, walls and floor materials, they are not definitive even in built phase, so in a future version we would give the possibility to the users to edit them. They could see and try the possible changes, based on the current project advancement and architectural rules (load-bearing walls etc.).

CONCLUSION AND PERSPECTIVES

This study case allows us to try and improve our methodology, by adapting a BIM model thanks to building project actors' needs, in this case the end users' needs. This VR prototype has to be tested by the architect and also in a project review meeting, to evaluate and improve it thanks to the user feedback. Notice that we are working currently on our BIM to VR automated preparation module: processing the BIM data, modifying the 3D model for a VR purpose thanks to these data etc. Finally, notice that we are thinking on linking that to a planned virtual walkthrough.

REFERENCES

1. H-J.Bullinger, W.Bauer, G.Wenzel, R.Blach. Towards user centred design (UCD) in archi- tecture based on immersive virtual environments, Computers in Industry, 61 (4), pp. 372- 379, 2010.
2. M.Heidari, E.Allameh, B.De Vries, H.Timmermans, J.Jessurun, F.Mozaffar. Smart-BIM virtual prototype implementation, Automation in Construction, 39, pp. 134-144, 2014.
3. A.Heydarian, E.Pantazis, E., A.Wang, D.Gerber, B.Becerik-Gerber. Towards user centered building design: Identifying end-user lighting preferences via immersive virtual environ- ments, Automation in Construction, 81, pp 56-66, 2017
4. A.Saleh, A.Rafi, P.Woods, X.Li, I.Hijazi, S.Cheng. Evaluation of three-dimensional com- puter visual materials to support user's participation in architectural design process, Journal of Intelligent and Fuzzy Systems -Volume 31, Issue 5, pp 2511-2523, 2016.
5. W.Wu, I.Kaushik, A BIM-based educational gaming prototype for undergraduate research and education in design for sustainable aging, Proceedings of the 2015 Winter Simulation Conference, pp 1091-1102, 2015.
6. W.Yan, C.Culp, R.Graf. Integrating BIM and gaming for real-time interactive architectural visualization, Automation in Construction, 20, pp 446-458, 2011.


PAPERS

L'INFORMATIQUE POUR LA VILLE INTELLIGENTE ET LE DEVELOPPEMENT DURABLE

JOSE TIBERIO HERNANDEZ PEÑALOSA,1 CARLOS JAIME BARRIOS HERNÁNDEZ2, MICHEL RIVEILL 3, YVES DENNEULIN4, CLAUDIA RONCANCIO4, FRÉDÉRIC MERIENNE5, HAROLD CASTRO1, ISMAEL PEÑA6, HELGA DUARTE6, OSCAR ALBERTO CARRILLO ROZO7, FRÉDÉRIC LE MOUËL8, ARTURO PLATA GOMEZ2, JORGE LUIS CHACÓN2, PABLO FIGUEROA1, RAUL RAMOS POLLÁN2 , GABRIEL RODRIGO PEDRAZA FERREIRA2 ET MIREILLE BIAY-FIALLO3.

1. Universidad de los Andes (UniAndes), Colombia
2. Universidad Industrial de Santander (UIS), Colombia
3. Université Nice-Sophia Antipolis (UNSA), France
4. Université de Grenoble (UDG), France
5. École Nationale des Arts et Meiières– Paris Tech (ENSAM- Paris Tech), France
6. Universidad Nacional de Colombia (UNAL), Colombia
7. École supérieure de Chimie, Physique, Électronique de Lyon (CPE-Lyon), France
8. Institut National des Sciences Appliquées de Lyon (INSA-Lyon), France

RÉSUMÉ

La collaboration franco-colombienne en technologies de l'information avancées, à partir la rechercher et le développement collaboratif, promit le déploiement et l'usage responsable du numérique pour bâtir des villes intelligentes et durables. Elle vise à intégrer des solutions technologiques pour améliorer la gestion urbaine, tout en respectant les ressources naturelles et les principes d'inclusion sociale. Cette approche encourage l'échange de données, la participation citoyenne et l'innovation ouverte. Elle cherche à réduire les inégalités numériques et à renforcer la résilience des territoires face aux défis climatiques et sociaux. Le numérique devient ainsi un levier stratégique pour un développement urbain éthique et collaboratif. Cet article montre les principaux enjeux vis-à-vis la formalisation de la collaboration pour des problématiques de ville durable.

1. INTRODUCTION

La mondialisation des échanges et l'organisation du travail que ce dernier induit, provoquent un flux migratoire important de la population rurale vers les villes. Ce mouvement mondial est encore plus critique dans les villes en développement rapide. La conséquence de cette situation est une énorme difficulté pour gérer de manière pertinente et cohérente la croissance des villes.

Des problèmes de natures diverses apparaissent : ceux-ci peuvent concerner directement chacun des citadins mais aussi en amont ceux qui sont en charge de la gestion de la cité. Nous pouvons citer par exemple, les problèmes liés à l'inadéquation des voies de circulation aux flux de voitures, l'augmentation de la pollution ou l'accès des citoyens aux services de la cité.

L'ensemble des partenaires du LIA sont engagés d'une manière ou d'une autre dans ce vaste chantier ou l'informatique (sous ses multiples facettes) peut permettre l'anticipation des problèmes en vue d'apporter des solutions à certains des problèmes de croissance rapide des villes ou d'offrir aux citoyens la possibilité d'être acteur du développement de leur cité. Ce projet de laboratoire international associé franco-colombien CATAI propose de fédérer les activités de recherche et de rendre plus cohérents les divers travaux menés par les uns et les autres, autour de l'apport des travaux menés par les uns et les autres sur l'apport de l'informatique à la ville intelligente avec pour cible de générer de la durabilité urbaine.

La force d'une telle collaboration est d'analyser des situations urbaines différentes avec leurs dynamiques propres en mettant en commun des concepts et outils numériques qui rendent réalistes le développement durable de nos villes.

2. CONTEXTE

Les difficultés induites par la croissance de la plupart des grandes zones urbaines sont fortement dépendantes du continent d'appartenance. En revanche, les confrontations entre villes de natures différentes de leurs problèmes et leurs démarches permettent de proposer des solutions originales.

Des jumelages et programmes communes entre des villes françaises et colombiennes existent (par exemple entre Grenoble et Bucaramanga) permettant d'échanger sur des problématiques communes.

Parmi ces problématiques communes, la gestion et l'accès aux données de la ville sont critiques car elles impliquent l'ensemble des acteurs de la ville (autorités, construction, maintenance, gestion, citadin). L'informatique est au cœur de ces problématiques.

Par ailleurs, il existe de nombreuses relations antérieures et actuelles dans le domaine de l'informatique entre la France et la Colombie. Ces relations sont nourries par le fait que de nombreux enseignants-chercheurs Colombiens ont effectué leur thèse en France.

Des travaux communs ont déjà eu lieu sur la thématique de l'informatique pour la ville intelligente. Côté français, 3 laboratoires sont impliqués : I3S, LIG et Le2i qui sont tous 3 des UMRs du CNRS dépendant de l'INS2i. Par ailleurs, aux vues de la thématique abordée, d'autres équipes de recherche pourront être associées aux travaux comme par exemple l'UMR ESPACE (CNRS / UNS) dépendant de l'INEE.

Les raisons évoquées motivent la proposition de la création d'un laboratoire international associé franco-colombien sur cette thématique. La constitution de ce laboratoire permettra de construire dans la durée un programme de recherche pour adresser certains verrous existants, apporter des solutions aux problèmes posés et diffuser ces solutions dans le milieu socio-économique au service du bien-être du citadin.

3. CONSORTIUM

3.1 PRÉSENTATION DE L'ÉQUIPE CATAI

L'équipe CATAI constituée par certains membres des laboratoires présentés ci-après comprendra des compétences complémentaires permettant d'adresser les problématiques de l'informatique au service de la ville intelligente. Par ailleurs, les plateformes technologiques des laboratoires partenaires seront mises au service du laboratoire CATAI. Les moyens suivants sont identifiés :

Domaine	Équipements
Calculateurs hautes performances	Infrastructures de Calcul de Grande Échelle Supercalculateurs (Grappes) Grilles de Calcul Infrastructures de Calcul de Haute Performance Embarquées
Visualisation interactive	Larges écrans de visualisation Salles d'immersion virtuelle Simulateurs de conduite
Centres de stockage et préservation des données	Infrastructures de stockage et de préservation des données spécialisés pour le projet. Serveurs dédiées de webservices et gateways spécialisés

3.2 PARTENAIRES

Les partenaires academiques et scientifiques impliquées dans la proposition, son les suiantes :

3.2.1 LABORATOIRE LE2I (FRANCE)

Le Laboratoire d'Electronique, d'Informatique et de l'Image (Le2i) est une Unité Mixte de Recherche (U.M.R. 6306) rattachée principalement à l'INSII et à l'INSIS en institut secondaire. Le laboratoire comprend 96 Enseignants-Chercheurs, 17 ingénieurs, techniciens et personnels d'encadrement administratif, et environ 95 doctorants et post-doctorants, le Le2i est le seul laboratoire en Sciences et Technologies de l'Information et des Communications (STIC) de la région Bourgogne. Il couvre un large spectre d'activités au sein de trois départements scientifiques :
A) Informatique,
B) Electronique et
C) Vision. Le Le2i est réparti sur 4 sites en Bourgogne (Dijon, Le Creusot, Auxerre et Chalon-sur-Saône).

Les Etablissements tutelles du Le2i sont :

- Université de Bourgogne
- Arts et Métiers ParisTech (ENSAM)
- CNRS

Le laboratoire Le2i développe des activités de recherche sur la thématique de la ville depuis de nombreuses années à travers l'équipe « réalité virtuelle » et le projet transversal « CheckSem ». L'équipe « réalité virtuelle » à l'institut image a été impliquée dans des projets de recherche mettant en œuvre l'immersion virtuelle au service du patrimoine bâti et du bâtiment (numérisation 3D, maquette virtuelle interactive, immersion virtuelle et réalité augmentée). Deux start-ups ont été essaimées en proximité avec ce champ d'application : la société On-Situ (créée en 2006) et la société Paztec (créée en 2013).

Le projet CheckSem s'intéresse à la modélisation sémantique et formelle de connaissances et leur manipulation pour le domaine du BiM sémantique. Une plateforme d'intelligence sémantique a été développée ainsi qu'une architecture orientée services autour d'un noyau combinant des triplestores, du graph mining, des métaheuristiques, et du model checking. Les travaux de recherche de Checksem ont permis en 2008 la création de la société

Active3D.

3.2.2 LABORATOIRE I3S (FRANCE)

Le laboratoire I3S est une unité mixte de recherche Université Nice Sophia Antipolis / CNRS associé à Inria. Il rassemble plus de 300 personnes dont 120 chercheurs sénior (enseignants-chercheurs de l'UNS ou chercheurs CNRS (18) et chercheurs Inria (12)) et 80 doctorants.

Le laboratoire I3S est au cœur d'un réseau de partenariats et de programmes collaboratifs, nationaux et internationaux, qui soutiennent et structurent son activité. Ainsi, il est actif dans plusieurs pôles de compétitivité, en particulier : SCS (Solutions Communicantes Sécurisées), PEGASE (Pôle aéronautique et spatial) et PASS (Parfums, Arômes, Senteurs et Saveurs).
Le laboratoire I3S développe depuis plusieurs années des travaux ayant un lien avec la thématique du LIA principalement autour des deux aspects suivants :

- Intelligence artificielle et principalement a) sur les systèmes muti-agent en particulier pour modéliser et simuler des comportements lorsque les agents ont des objectifs (individuels ou non) et des croyances sur le monde dans lequel ils évoluent, ce qui pourrait influencer l'ordre voir même le choix de leurs actions.Ces agents peuvent revoir leurs croyances en fonction de nouvelles informations. Le fait d'utiliser de tels agents dotés d'une représentation de leur état mental permet de modéliser de façon plus fidèle à la réalité les systèmes complexes où les acteurs sont des êtres humains. b) à la fouille de données pour proposer de nouvelles solutions par l'analyse des données passées sur l'évolution de villes
- Génie logiciel et principalement sur les techniques de développement orientées services à base de composants permettant d'augmenter la productivité et la réutilisabilité en découplant les différentes facettes de logiciels. L'ingénierie dirigée par les modèles permet de capitaliser le savoir-L'ingénierie dirigée par les modèles permet de capitaliser le savoir- faire dans les modèles loin des plate-formes technologiques tout en maintenant la traçabilité avec des applications en cours d'exécution.
- Les problèmes relatifs au passage à l'échelle en terme de stockage et d'accès aux données, de performances dans les calculs et dans les propagations d'information sont particulièrement critiques dans le cadre des applications ciblées. Ainsi les infrastructures à grande échelle deviennent un support indispensable et transparent pour l'exécution d'applications complexes nécessitant un très grand nombre de ressources distribuées à travers des principes de logiciels divers comme les services, le cloud computing et de grid computing, reposant dans le contexte des villes sur des supports très hétérogènes aux performances diverses.

Sera associé au laboratoire I3S, une équipe de recherche du l'UMR ESPACE (4 enseignants-chercheurs - 1 chercheur CNRS) qui développe des travaux autour des interactions multi-échelles et fonctionnement des systèmes territoriaux. L'approfondissement de la connaissance des systèmes territoriaux, et plus spécialement des interrelations espace-environnement-société, constitue l'une des finalités du projet de recherche.

3.2.3 LABORATOIRE LIG (FRANCE)

Le Laboratoire d'Informatique de Grenoble rassemble près de 500 chercheurs, enseignants-chercheurs, doctorants et personnels en support à la recherche. Ils relèvent des différents organismes et sont répartis sur les deux sites du LIG : le campus de Grenoble et Montbonnot.

Le projet scientifique du LIG est l'"Informatique ambiante et durable". L'ambition est de s'appuyer sur la complémentarité et la qualité reconnue des 22 équipes de recherche du LIG pour contribuer au

développement des aspects fondamentaux de la discipline (modèles, langages, méthodes, algorithmes) et pour développer une synergie entre les défis conceptuels, technologiques et sociétaux associés à cette thématique que l'on retrouve très présente dans le projet CATAI.

Le LIG contribuera à CATAI sur les aspects gestion de données à large échelle, calcul intensif et mouvements de données associés ainsi que sur le traitement des informations spatio-temporelles, depuis leur acquisition au moyen de capteurs physiques ou citoyens ou par le biais de sources de données officielles, ou bien encore liées et ouvertes, jusqu'à leur restitution sous diverses formes : cartes, graphes et graphiques dynamiques et interactifs, jeux de données structurés ou non structurés, services web géographiques, *etc.*, sans omettre l'association de métadonnées renseignant sur la provenance de ces données et sur leur qualité intrinsèque au regard d'un certain nombre de critères définis.

3.2.4 LABORATOIRE IMAGINE/COMIT-UNI ANDES (COLOMBIE)

Les équipes de recherche IMAGINE (Informatique visuelle) et COMIT (Communications et Technologies de l´information), sont rattachées au département d´Ingénierie Informatique (Sistemas y Computación) de l´Université de los Andes, â Bogotá. L´équipe IMAGINE comprend 4 Enseignants-Chercheurs, et 7 doctorants et post-doctorants. L´équipe COMIT comprend 9 Enseignants-Chercheurs, et 12 doctorants et post-doctorants.

Le laboratoire Imagine/Comit développe des activités de recherche sur la thématique de la ville depuis de nombreuses années â travers des projets de visual analytics et bigdata analysis Des collaborations avec des équipes de génie de transport, et d´urbanisme ont données lieu â des projets appliqués â la ville de Bogota.

Cette dynamique a permis la création de nouvelles sociétés, particulièrement la société Datatraffic (créée en 2007) qui offre des services de valeur ajouté supportés par une cartographie digitale augmentée Le projet TaCAT permet la mise en œuvre de nouveaux outils interactifs pour la prise des décisions dans le domaine des systèmes urbains. Des particularités des sources d'information, quantité et qualité des données, différents outils d'analyse et différents stakeholders pour la prise de décisions, font partie des caractéristiques de ce projets.

IMAGINE/COMIT va contribuer à CATAÏ dans les domaines de bigdata, Visual Analytics, Internet of things, Realité virtuelle et IHM notamment. L'analyse des données spatio temporelles, les nouveaux environnements interactifs d'analyse collaboratif, l´integration des données heterogenes et des outils d'analyse HPC nous offrent des opportunités exceptionnelles de collaboration scientifique.

3.2.4 CENTRE DE CALCUL D'HAUT PERFORMANCE ET SCIENTIFIQUE DE L'UNIVERSITÉ INDUSTRIELLE DE SANTANDER (SC3UIS) (COLOMBIE)

Le Centre de Calcul d'Haut Performance et Scientifique a été crée en 2011 à l'Université Industrielle de Santander dans la cadre du projet « Technologies avancées de l'information et la communication pour la science et la génie de l'orient colombien », Le processus de création du centre à compromis un accompagnement de côte français vis à vis le Laboratoire d'Informatique de Grenoble, l'INRIA Rhône Alpes

y le Laboratoire d'Informatique Signaux et Systèmes de Sophia Antipolis. La collaboration des laboratoires et institutions français à permis la définition des lignes de travaux de recherche et développement, une planning d'évolution du centre à cinq ans de façon opérative y administratif et la planification des lignes de formation articulés avec les programmes d'école d'ingénieur, master, doctorat en informatique et formation continue.

Le centre SC3UIS a comme objectifs le support de projets de recherche, développement et innovation dans les moyennes academique et industrielle. Le support involucre des activites de formation dans different niveaux (utilisateurs scientifiques, developpeurs, administrateurs d'infrastructure, etc) et aussi la médiation strategie entre entreprises, les institutions de gouvernement et le secteur scientifique. Son activité il n'est pas uniquement addressé aux problematiques regionaux, sinon colombiennes et en projets internationaux, principalement en amerique latine.

Le centre SC3UIS au niveau scientifique a deux importants composants: un qui garanti une activitié autour des sciences de la communication et l'information avec des lignes de recherche en architectures de grande echelle, analytique des données

- donc le big data, l'architecture et génié de logiciel, le calcul de haute performance, la programmation concurrente et parallele, des systemes distribués et les systèmes qui supportent le processus parallele. L'autre s'occupe des applications scientifiques et la recherche au tour des problèmes en science et ingenierie qu'on besoin de calcul de haute performance: l'eau et l'environnement, l'astronomie et l'astrophysique, les sciences de la vie, les sciences sociaux et humaines, l'energie, l'eau et le gaz, le traitement des données physique-chimie et les sciences de la terre.

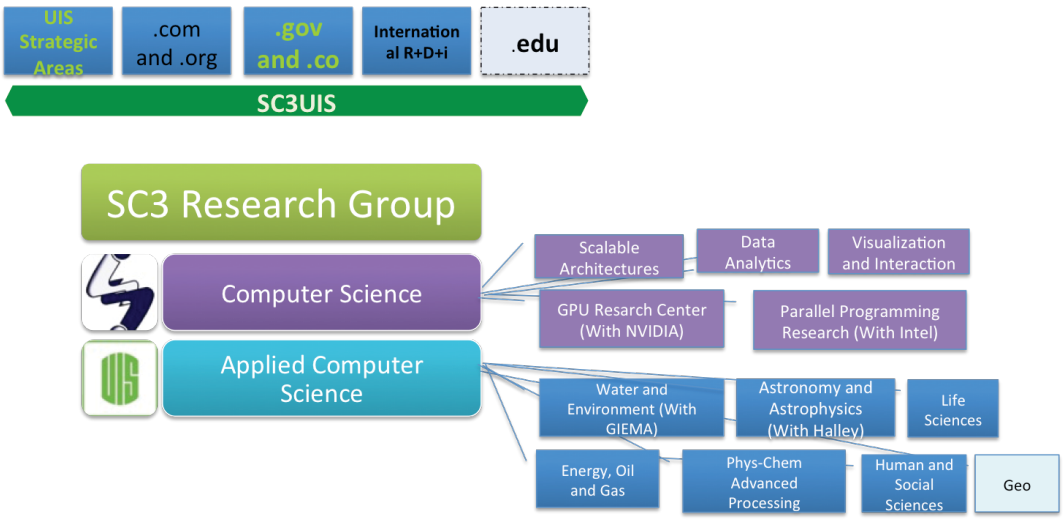


Figure 1. organisation de l'activité scientifique a supprime : X

3.2.6 GRUPO DE INVESTIGACIÓN GEOURBE DE LA UNIVERSIDAD NACIONAL DE COLOMBIA (COLOMBIE)

Le groupe de recherche GEOURBE regroupe des chercheurs intéréssés par les thématiques urbaines et régionales, à partir d'une approche géopolitique basée sur l'idée selon laquelle la ville, et par extension toute unité territoriale, est un produit social résultant d'intérêts et de valeurs sociales conflictuelles.

Le groupe se compose essentiellement de géographes, mais il est ouvert à d'autres disciplines reliées à la géographie par l'intérêt dans la dimension spatiale des processus sociaux. Il est inscrit dans la Red Internacional de Fuentes de Información y Conocimiento para Gestión de Ciencia, Tecnología e Innovación (SCienTI)de COLCIENCIAS.

3.2.7 LABORATOIRE "ETUDE DES STRUCTURES, DES PROCESSUS D'ADAPTATION ET DES CHANGEMENTS DE L'ESPACE" (ESPACE) - UMR 7300 CNRS

ESPACE associe le CNRS et des laboratoires de trois universités : Nice Sophia-Antipolis, Aix Marseille Université et Avignon. Antipolis, Aix Marseille Université et Avignon. Il est rattaché à deux instituts du CNRS, l'Institut National des Sciences Humaines et Sociales et l'Institut National Ecologie et Environnement. L'UMR ESPACE dirige l'Observatoire Homme-Milieu Littoral méditerranéen et appartient au Laboratoire d'Excellence « Dispositif de Recherche Interdisciplinaire pour les Interactions Hommes-Milieux » (DRIIHM). ESPACE-Nice rassemble 16 enseignants-chercheurs et chercheurs CNRS, 9 ingénieurs et techniciens et 17 doctorants.

Les travaux du laboratoire ESPACE sont principalement consacrés aux espaces urbanisés plus ou moins denses, à l'analyse de leur organisation, de leur évolution, de leurs interactions avec les systèmes naturels et agricoles avec lesquels ils sont en contact. Ces espaces en mutation constante sont étudiés selon une approche systémique qui permet d'appréhender de manière globale les divers champs d'étude, mais aussi d'individualiser les éléments et les facteurs qui semblent pertinents pour étudier cette complexité.

Les phénomènes urbains sont analysés dans leur dimension multi-niveaux afin de déceler les types de relations inter-scalaires et déterminer le rôle des jeux d'échelles dans le fonctionnement des systèmes territoriaux. Les travaux portent sur la caractérisation des géosystèmes urbains et leur trajectoire déclinée selon deux concepts : le déploiement (conservation de l'équilibre) et l'évolution (transition d'un état à un autre).

Utilisant les outils d'analyse spatiale et les SIG, le laboratoire s'est très tôt intéressé à la modélisation t à la simulation dans un but heuristique : il fut notamment leader sur les approches individus-centrées en géographie dès 1998 avec un programme de recherche où des modèles de simulation de la mobilité résidentielle ont été réalisés à partir d'approches utilisant les automates cellulaires et les systèmes multi-agents. agents.

La géoprospective, c'est-à-dire l'anticipation des changements spatiaux par différents types de modélisation et de simulation, est une des spécialités du laboratoire.

4. PROJET SCIENTIFIQUE

4.1 CONTEXTE ET ENJEUX

Le projet scientifique proposé consiste à faire collaborer des unités de recherche ayant des compétences complémentaires dans divers domaines de l'informatique au service de la ville intelligente et la soustenibilité, centrées dans les besoins humains. La ville de demain aura des besoins cruciaux dans la

constitution, la gestion et l'accès à ses données. Ainsi, la ville doit avoir son modèle numérique évoluant avec elle. Les données collectées peuvent être apportées par des sources diverses et très hétérogènes au niveau de leurs fiabilités.

La participation active des citadins permettra de mettre le citoyen au cœur de la ville tant pour informer que pour s'informer et co-décider. Les services induits par cette ville inter-connectée pourront être optimisés (service à la personne, sécurité, mobilité, accès aux infrastructures, guide touristique…). Les co-décideurs de leur environnement urbain auront accès à des systèmes de visualisation interactive pour l'état des lieux et l'état projeté (simulation) des mobilités (temps de trajets), pollution, économie des quartiers

Les enjeux de ces besoins sont au carrefour de différents métiers de l'informatique : la constitution de base de données avec traçabilité, la gestion de masses de données lourdes, hétérogènes et multi-échelles, la visualisation interactive de données complexe.

4.2 POSITIONNEMENT

Sur la thématique de la ville du futur, plusieurs universités prestigieuses développent des programmes de recherche ; preuve que les enjeux sociétaux de la ville sont prégnants dans le monde. Le tableau suivant synthétise quelques laboratoires (tous relativement jeunes) existants dans le domaine. La liste n'est bien sûr pas exhaustive.

laboratoire	Objectifs
City Science Initiative MiT Media Lab (USA) http://cities.media.mit.edu/	To build the cities that the world needs, we need a scientific understanding of cities that considers our built environments and the people who inhabit them. Six initial themes represent a cross section of the interdisciplinary research that will be undertaken to address the major challenges associated with global urbanization. Urban analytics and modeling Incentives and governance Mobility networks Places of living and work Electronics and social networks Energy networks
Smart City Lab University of Bologna (Italy) http://smartcity.csr.unibo.it/	Research in urban ict innovation for a better life. The "Smart City" model is strictly related to the ICT infrastructures that the city itself shares with the citizens; these innovative architectures provide a more efficient conservation of the fixed heritage and a better standard of living, thanks to modern and full-accessible services.
Smart City Institute Accenture, Belfius, Ville de Liège, Université de Lière (Belgique) http://labos.ulg.ac.be/smart-city/homepage/	The researchers of the Smart City Institute work on the development of a scientific expertise – with an international perspective and impact – on the management of smart cities.

City Lab Inria (France) https://citylab.inria.fr/	The Inria Lab CityLab@Inria studies ICT solutions toward smart cities that promote both social and environmental sustainability. A strong emphasis of the Lab is on the undertaking of a multi-disciplinary research program through the integration of relevant scientific and technology studies, from sensing up to analytics and advanced applications, so as to actually enact the foreseen smart city Systems of Systems.
Smart cities research Center Univ of California Berkeley (USA) http://smartcities.berkeley.e du/	The Research Center was established at UC Berkeley as a part of inter-disciplinary effort to advance quantitative modelling of urban systems. It's core is at CEE Systems and Transportation. We work on mathematical models and data analytics with public agencies as well as private companies and focusing on fundamental research in a broad area of Smarter Cities.
Smart Cities Center Columbia University (USA) http://datascience.columbia.edu/smart-cities	Research conducted by the Smart Cities Center develops and monitors sustainable urban infrastructure and buildings, improves the power supply through smart grid technology, detects and counteracts problems with aging urban infrastructure, calculates and communicates optimal transportation routes under congested traffic conditions, and deploys ubiquitous sensing devices to facilitate everyday activities in a crowded urban environment.
Laboratoire	Objectifs
City Science Initiative MiT Media Lab (USA) http://cities.media.mit.edu/	To build the cities that the world needs, we need a scientific understanding of cities that considers our built environments and the people who inhabit them. Six initial themes represent a cross section of the interdisciplinary research that will be undertaken to address the major challenges associated with global urbanization. Urban analytics and modeling Incentives and governance Mobility networks Places of living and work Electronics and social networks Energy networks
Smart City Lab University of Bologna (Italy) http://smartcity.csr.unibo.it/	Research in urban ict innovation for a better life. The "Smart City" model is strictly related to the ICT infrastructures that the city itself shares with the citizens; these innovative architectures provide a more efficient conservation of the fixed heritage and a better standard of living, thanks to modern and full- accessible services.
Smart City Institute Accenture, Belfius, Ville de Liège, Université de Lière (Belgique) http://labos.ulg.ac.be/smart-city/homepage/	The researchers of the Smart City Institute work on the development of a scientific expertise – with an international perspective and impact – on the management of smart cities.
City Lab Inria (France) https://citylab.inria.fr/	The Inria Lab CityLab@Inria studies ICT solutions toward smart cities that promote both social and environmental sustainability. A strong emphasis of the Lab is on the undertaking of a multi- disciplinary research program through the integration of relevant scientific and technology studies, from sensing up to analytics and advanced applications, so as to actually enact the foreseen smart city Systems of Systems.

Smart cities research Center Univ of California Berkeley (USA) http://smartcities.berkeley.e du/	The Research Center was established at UC Berkeley as a part of inter-disciplinary effort to advance quantitative modelling of urban systems. It's core is at CEE Systems and Transportation. We work on mathematical models and data analytics with public agencies as well as private companies and focusing on fundamental research in a broad area of Smarter Cities.
Smart Cities Center Columbia University (USA) http:// datascience.columbia .edu/smart-cities	Research conducted by the Smart Cities Center develops and monitors sustainable urban infrastructure and buildings, improves the power supply through smart grid technology, detects and counteracts problems with aging urban infrastructure, calculates and communicates optimal transportation routes under congested traffic conditions, and deploys ubiquitous sensing devices to facilitate everyday activities in a crowded urban environment.
CASA (The Centre for Advanced Spatial Analysis) UK. http://www.bartlett.ucl.ac.uk /casa	CASA is engaged in generating new knowledge and insights for use in city planning, policy and design and drawing on the latest geospatial methods and ideas in computer-based visualisation and modelling.

STRATÉGIE SCIENTIFIQUE

La stratégie scientifique du laboratoire CATAÏ aura pour contexte les aspects liés à l'évolution rapide des villes soit pour offrir des outils de planification ou de simulation aux décideurs, soit pour offrir des outils participatifs aux citoyens.

En effet, les outils informatiques peuvent s'avérer particulièrement utile dans quatre thématiques : le transport, la santé publique, les réseaux de communication, la gestion des risques.

L'originalité de la proposition réside dans la complémentarité des partenaires pour adresser cette problématique tout en ayant un socle commun de compétences. Compte tenu des enjeux identifiés, le laboratoire CATAI focalisera son action sur la chaîne de la valeur depuis la constitution des données de la ville, leur gestion jusqu'à leur visualisation interactive.

Compte tenu du caractère applicatif de la proposition, des compétences en génie logiciel seront mis en œuvre de façon à proposer des démonstrateurs technologiques.

Ainsi, les compétences des différents partenaires seront mutualisées pour adresser 4 champs permettant de couvrir la chaîne de valeur identifiée :

- Visualisation interactive
- Modèle multi-échelle
- Calcul de haute performance
- Génie Logiciel

Par ailleurs, les compétences des partenaires impliqués dans le laboratoire CATAI ne seront pas suffisantes pour embrasser l'ensemble des problématiques scientifiques envisagées qui induisent une approche fortement pluri-disciplinaire. Aussi, les travaux seront menés en partenariat avec des laboratoires spécialisés dans l'étude de ces domaines applicatifs (géographie, urbanisme, architecture, histoire, sociologie, ergonomie, psychologie cognitive…).

4.1 AXES THÉMATIQUES

Les thèmes identifiés correspondent aux 4 axes thématiques résumés dans la figure suivante. Les dimensions multi-modèles et génie logiciel sont transverses à l'ensemble du projet et donc les interfaces entre le calcul haute performances et la visualisation.

Axe 1 – Modélisation multi-échelle

L'objectif de cet axe de recherche est d'aboutir à des modélisations de la ville permettant de maîtriser la complexité de l'existant et de son évolution et de proposer des services innovants et des connaissances utiles à la prise de décision en matière de politiques urbaines. Deux thèmes font l'objet des modélisations dans cet axe : la modélisation des données et des connaissances et la modélisation spatio-temporelle des formes et des processus urbains.

Sous-axe 1. Modélisation de données et des connaissances sur la ville

D'un point de vue de la gestion des données, l'enjeu scientifique est de définir des processus de migration de sources de données traditionnelles (systèmes d'information classiques, données structurées) pour qu'elles eviennent exploitables dans une logique ouverte au service de la ville et des citoyens. L'objectif est de pouvoir tirer profit de telles sources conjointement avec des données actuelles produites et stockées avec des technologies récentes (systèmes NoSQL, approches à base de connaissance). Il y aura à maitriser une masse de données combinant données structurées, non structurées, textuelles et une dimension spatio-temporelle presque omniprésente. Ces caractéristiques, qu'on qualifierait de big data, sont accentuées d'autant plus par la dimension ubiquitaire des systèmes déployés au service des villes intelligentes.

Les recherches concernent la maîtrise de la collecte et l'exploitation de données relevées par des capteurs physiques ou logiciel, couplés à des stratégies où l'on contrôle fortement la production des données ou au contraire des approches de type crowdsourcing basée sur une participation très subjective, et quasi opportuniste des habitants.

Diverses échelles de collecte et de traitement devront être possibles afin de répondre aux besoins et aux contraintes de coût et qualité des utilisateurs. Les citoyens-capteurs ou produsers, à l'image des processus de démocratie participative, pourront être sollicités pour des campagnes de collectes de données ou encore des enquêtes d'opinion.

Ces sondages nécessitent des infrastructures de données capables : i) en amont, de sélectionner au besoin sur critères divers – degré d'expertise, localisation, âge, profession, etc. – tout ou partie d'une population, de qualifier les données ainsi collectées, de construire ou compléter les métadonnées indispensables, de lier ces données et métadonnées avec des données institutionnelles, éventuellement de les assembler en des séries temporelles longues et surtout cohérentes ; ii) en aval de les restituer à la fois à des experts pour que soient menées des analyses conduisant à la prise de décision dans les domaines de l'urbanisme, de l'aménagement du territoire, de la prévention des risque, de la santé etc., mais également aux citoyens, partie prenante dans ce processus de production et d'exploitation des données de la ville, à des fins d'information. Si les technologies actuelles, notamment celles préconisées par l'Open Geospatial Consortium, permettent, en principe, la construction de telles infrastructures, la mise en place d'un tel flux de données combinant données institutionnelles et citoyennes, reste largement un défi à relever.

D'un point de vue des systèmes d'information, comprendre et modéliser des socio- écosystèmes complexes reste un problème non maitrisé. Les systèmes « supportant » les villes intelligentes sont

intrinsèquement complexes dans le sens où ils sont très dynamiques et combinent un éventail très large de services et d'acteurs appartenant à des organisations hétérogènes et autonomes. Les approches de conception de systèmes socio-techniques complexes constituent également un sujet de recherche à explorer.

Sous-axe 2. Modélisation des formes et des processus urbains de changement rapide

Les transformations rapides des espaces urbanisés dans le monde exigent une planification capable de faire face aux défis économiques, sociaux et environnementaux de la ville d'aujourd'hui. Cette planification doit être adaptée à la complexité des phénomènes urbains, mais aussi au rythme des évolutions qui s'imbriquent dans différentes échelles spatiales (des quartiers aux réseaux de villes) et temporelles (de la mobilité quotidienne aux transitions urbaines et métropolitaines).

Les recherches actuelles en modélisation urbaine tendent à se focaliser sur le phénomène d'extension des aires urbaines et sur les conséquences qui en résultent, notamment sur le plan environnemental. Or, dans nombre de pays du Sud, et dans une moindre mesure, du Nord, le processus de croissance urbaine se manifeste également par des transformations de la structure urbaine qui touchent les centres comme les périphéries déjà urbanisées. Ce réaménagement continuel ne relève pas uniquement d'opérations d'urbanisme. Des transformations spontanées, en lien ou non avec les actions programmées, pouvant se produire à des rythmes très rapides, transforment la ville dans ses dimensions horizontales et verticales comme dans ses fonctions. Dès lors, l'anticipation de la régénération intra-urbaine et de l'étalement urbain, par la modélisation et la simulation, est d'un intérêt majeur.

La modélisation urbaine actuelle se limite souvent, soit à la simulation des mobilités quotidiennes en fonction de la répartition des localisations des populations et des fonctions urbaines dans des modèles de type LUTI (Land Use/Transport Interaction Models), soit à la simulation de l'évolution de l'occupation du sol par télédétection et automates cellulaires pour étudier l'extension de la tache urbaine. Or, il est rare de concevoir des modèles qui offrent la possibilité de faire le lien entre les différentes échelles spatiales et temporelles et de formaliser le lien entre les formes et le fonctionnement de la ville. La modélisation urbaine se voit donc confrontée à plusieurs difficultés.

Une première difficulté provient de la nécessité de définir les échelles appropriées à la description des différents processus urbains. Les phénomènes de métropolisation doivent être analysés à différentes échelles spatiales, du quartier au réseau de villes. De façon analogue, le rôle des technologies de l'information et de la communication dans la mobilité quotidienne et de la complexité de l'organisation spatiale des territoires urbains impose une imbrication des échelles temporelles dans les analyses.

La deuxième difficulté provient de la relation entre les différentes échelles. La modélisation cherche souvent à décrire l'auto-organisation de la ville par l'interaction entre ses éléments constitutifs dans une approche « bottom-up ». Or, il est indéniable qu'il est indispensable d'intégrer dans cette modélisation le rôle des politiques publiques, des agents économiques et des facteurs externes (modèle de développement économique, caractéristiques physiques et géographiques, histoire urbaine, etc.) sur le comportement du système correspondant à une approche « top- down ».

La troisième difficulté est celle de l'intégration des flux grandissants de données dans la calibration et la validation de modèles de plus en plus complexes. La modélisation informatique se voit aussi confrontée à la nécessité d'apporter des réponses en temps réel permettant d'ajuster les infrastructures urbaines au comportement changeant de la ville. La disponibilité d'informations a transformé la manière dont les modèles sont conçus et leur rôle dans la prise de décision par les décideurs publics. Ils jouent un rôle

important aussi dans la représentation et la communication des phénomènes. Ils se doivent donc d'assurer leur intelligibilité auprès des différents acteurs de la ville, ce qui impose de nouveaux défis dans la visualisation informatique.

Les objectifs des différents types de modélisation de l'évolution urbaine dans le cadre de ce projet sont :

- décrire les trajectoires des dynamiques urbaines en intégrant la croissante complexité spatio-temporelle de leur fonctionnement, notamment dans le contexte des fast changing cities colombiennes
- mettre en relation les formes urbaines et le fonctionnement de la ville en termes de mobilité (quotidienne et résidentielle), organisation socio-spatiale et impact environnemental
- fournir à la population civile et aux décideurs des outils, notamment cartographiques, pour anticiper les changements dans un contexte d'évolution rapide des formes urbaines
- identifier la possible émergence de phénomènes nouveaux à travers l'analyse de signaux faibles voire contradictoires dans l'évolution de la ville
- étudier les transformations des formes urbaines et pas seulement la croissance de la tache urbaine à partir de simulations à base d'agents morphologiques
- analyser les formes urbaines, non seulement à partir de leur extension surfacique, mais aussi dans les relations allotopiques induites par les réseaux de circulation et dans leur composante verticale à travers des modélisations en 3D
- intégrer des formalismes à base d'incertitude pour faire face à des données manquantes ou multi-sources, aux problèmes d'appréhension d'objets et de schémas spatiaux et à la subjectivité inhérente à la représentation des phénomènes urbains
- développer des passerelles (protocoles de formalisation, modélisation des connaissances, intégration aux règles de la géo-simulation, validation des modèles, etc.) entre les approches modélisatrices à la ville et les connaissances expertes des villes colombiennes produites par des approches plus classiques en géographie et en urbanisme.
- Développer des infrastructures de données spatiales et temporelles, capables de stocker et de restituer ces données dans toute leur diversité, en maîtrisant leur imperfection, en améliorant continuellement les processus de production et de collecte, associés, et, de là, d'en garantir une utilisation avertie.

Axe 2 – Calcul de haute performance

Le modèle ainsi constitué sert de base de travail pour réaliser des simulations et de traitement des donnes masives permettant de représenter ce qui ne se voit pas forcément rapidement ou de réaliser un état projeté (simulation d'indicateurs en fonction d'hypothèses d'aménagement urbain). Des simulations avec de nombreux couplages seront réalisées. La complexité des données et le nombre de variables induiront des verrous scientifiques liés au calcul haute performance sur de grandes quantités de données dont le stockage et l'accès est aussi un enjeu. La nature incertaine de certaines de ces données pourra également faire l'objet de proposition de méthodes de calculs dédiées. L'évolution de la ville pourra ainsi être projetée et permettra d'aider les co-décideurs dans l'aménagement urbain. Des verrous importants sont également à prévoir pour faire le nécessaire dialogue des méthodes et outils propres à chaque métier intervenant dans le changement de la ville (architecte, urbaniste, ingénieur).

Sous-axe 2.1 - : Infrastructures heterogenes de calcul de haute performance et efficience energetique de calcul:

Le support technologique pour l'execution des modeles de simulation et des applications qui permettent l'execution des logiciels pour le traitement des grandes volumes des donnees, posent des differents

questions associés avec le performance, l'efficience en la consommation energetique des processus (vis à vis l'exascale), la portabilitié et la possiblité de utiliser des platforms heterogenes pour accelerer les calculs ou les faire de facon embarqué (traitement des données in situ), toujours en garantir le haute performance.

L'utilisation de differents infrastructures suivent des importants variations dans la construction des algorithms, les mecanismes d'implementation des algorithms, les langages de programmation et l'evaluation de performance. Dans une autre cote, plus relié à l'architecture materielle et sa liaison avec le logiciel, il existe des particularités au niveau de compilation, de intergiciel et de systemes d'exploitation.

Sous-axe 2.2 - Architectures de grande echelle:

Les architectures de grande echelle impliquant des composants logiciels et materiels pour le traitement intensif et distribué des donnees, sur des platforms grid ou cloud. Les problematiques associées, comme la gestion de l'information, la securité, l'interoperabilité entres autres, permettent ne seulement l'interaction sinon le passage à l'echelle des problèmatiques et des données.

Le modèle de visibilité de cloud (IaaS, PaaS et Saas) presentent de façon principale l'interaction technologie-humain et la rélation données-implementation, associées aux problèmes technologiques (langages de programmation, observation de performance, coherence des données et de procées, tolerance aux fauts, etc), que sont dérivées de l'utilisation de platforms de grande echelle.

Axe 3 – Visualisation interactive

L'axe 3 s'intéressera plus spécifiquement aux interfaces permettant aux acteurs de la ville un accès aux données. Une maquette virtuelle de la ville et un environnement de cartographie dynamique basé sur les principes de la géovisualiastion, constitueront ainsi des outils au service de la collaboration entre urbaniste, citoyen, ingénieur, architecte, décideur. Par ailleurs, la conception et mise en œuvre dénvironnements interactifs de géovisualisation pour l'analyse exploratoire et l'aide à la décision jouera un rôle d'intégrateur entre les partenaires dans les actions du LIA. Cet axe de recherche est structuré en 5 sous-axes suivants, mais aussi il est très attaché à l'axe 2, du sur le besoin de visualisation remote pour des ambiances de collaboration, avec la possibilité d'un calcul et traitement des donnés in situ ou viceverse. La grande echelle alors involucre la proposition des services assocités au haut performance, comme les niveaux des services cloud et la visualisation interactive avancée et le travail collaboratif exprimes dans des sections suivantes.

Sous-axe 3.1 - Extraction de maquette virtuelle :

A partir des modèles conçus dans l'axe 1, une maquette virtuelle doit être construite de façon à permettre une visualisation interactive de la ville intégrant les données. Compte tenu du caractère interactif de la visualisation, les données doivent être simplifiées tout en possédant les éléments justes nécessaires à leur utilisation. La problématique de ce sous-axe est donc liée à l'extraction de maquettes virtuelles à partir des modèles multi-échelles et exhaustifs représentant la ville. Il s'agit par conséquent de développer des méthodes et outils permettant l'adaptation des données à l'application et aux utilisateurs de l'application. Pour mener à bien cette adaptation des données, des critères seront recherchés en liaison avec le profil de l'utilisateur ou des utilisateurs de la maquette virtuelle créée et avec les caractéristiques de l'application (visite virtuelle, revue de projet…). Par exemple, le profil de l'utilisateur pourra guider des besoins en calculs (par exemple de flux, de distances, d'hypothèses urbanistiques) qui seront par la suite proposés à l'utilisateur dans sa session de visualisation avancée.

Sous-axe 3.2 - Métaphores de visualisation :

Les données nécessaires à la compréhension de la ville sont plurielles et complexes et ne se limitent pas à sa géométrie. Il peut s'agir de résultats de calculs (liés à la géométrie) ou des hypothèses ou modifications urbaines. Il peut également s'agir d'informations sur la qualité des données (en termes d'incertitude ou de qualification du fournisseur de la donnée).

Ce sous-axe s'attachera à proposer des techniques de visualisation et plus particluièrement de géovisualisation permettant de représenter une variété possibles de données et d'indicateurs en fonction du contexte d'utilisation, des objectifs attendus (communication, analyse, exploration) et des utilisateurs finaux (publics, experts …), ainsi que des différents types de modélisation urbaines.

Des techniques existent dans la littérature: multifenêtrage synchronisé intégrant des représentations cartographiques, graphiques et/ou multimédia; cartographie dynamique et animée pour la représentation des dynamiques des territoires, cube spatio-temporel pour la représentation des mobilités urbaines.

Toutefois, face à la densité de données à représenter et à leur hétérogénéité, il est nécessaire de les faire évoluer, tant sur le plan des variables visuelles (couleur, forme, transparence, étiquettes de valeur ...) et des variables dynamiques (clignotement, déplacement, apparition, durée …) habituellement utilisées que sur celui des modes d'expression visuelles.

Par ailleurs, la combinaison des modes et des techniques de visualisation au sein d'un même environnement n'est pas aisée à mettre en œuvre pour satisfaire aux exigences de l'utilisateur et à la complexité des données.

Sous-axe 3.3 - Techniques de navigation et interaction :

Les données peuvent être très importantes pour ce qui concerne leurs quantités, ce qui suppose la possibilité de naviguer dans l'espace virtuel des données. Des techniques de navigation seront explorées et évaluées dans le contexte de la ville.

En fonction du profil de l'utilisateur, de l'application et du système de visualisation, la navigation pourra se faire avec une vue en miniature de la ville (survol de la ville) ou vue en position de piéton dans la ville.

Les techniques de navigation pourront être la technique de la carte virtuelle (avatar google map), la technique du monde en miniature, la technique en navigation à la première personne ou toute autre technique existante dans la littérature.

Une technique de navigation mal maîtrisée sur le sujet peut induire le mal de simulateur. Les techniques de navigation sont gérées par le mapping entre les mouvements de l'utilisateur (doigts sur un Joystick par exemple ou mouvement plus naturel) et les profils de vitesse et accélération de la navigation. Les effets de la technique de navigation et du mapping sur la navigation du sujet pourront faire l'objet de travaux de recherche.

Lors de l'expérience de visualisation interactive, l'utilisateur sera également amené à interagir avec les données. Les modalités d'accès aux données pertinentes pour son utilisation en fonction de son profil constituent un verrou important.

Les façons d'adresser les requêtes sur les données et d'accéder aux données demandées seront étudiées.

De la même façon, lors d'une session de travail collaboratif sur les données, il serait intéressant de proposer aux utilisateurs des moyens d'annoter les données pour conserver une mémoire du travail de collaborations.

Les méthodes et outils d'annotation et la remontée des annotations vers les données natives pourront faire l'objet d'études particulières.

Sous-axe 3.4 – Techniques de visualisation
interactive avancée :

En fonction de l'application, différents systèmes de visualisation interactive pourront être mis en œuvre. On pense aux écrans immersifs permettant une visualisation de la maquette numérique à grande échelle.

L'utilisateur peut être immergé dans la maquette virtuelle de la ville à l'échelle 1 et partager son expérience avec d'autres utilisateurs. Des systèmes collaboratifs plus simples de type table tactile pourront être mis en œuvre pour des applications particulières également.

Ces deux dispositifs sont utilisés dans un contexte de type laboratoire ou bureaux d'études. Ils permettent une bonne immersion avec les données grâce à leurs grands écrans. Un autre type de dispositif permettant la visualisation sur le site de la ville sera étudié.

Il s'agit de la réalité augmentée qui permet une interaction in situ avec les données numériques. La problématique est de permettre une adéquation entre les données et le site réel visualisé.

Sous-axe 3.5 - Travail collaboratif :

Un intérêt majeur de ces dispositifs complexes de visualisation interactive avancée réside dans la possibilité de mettre plusieurs utilisateurs de profils différents en collaboration avec les données.

Les problématiques posées par ce verrou sont relatives à l'accès aux données lors d'une visualisation sur site par la réalité augmentée ou lors de visualisation cartographique, l'échange de données entre dispositifs distants ainsi que l'adaptation des données en fonction du dispositif dans le cas de collaboration distantes entre systèmes asymétriques. Des protocoles de collaborations devront être définis au préalable.

Axe 4 – Génie logiciel

L'axe 4 concerne le génie logiciel. Cet axe est transversal aux 3 axes précédents et sert de colonne vertébrale aux actions du laboratoire CATAI. Il s'agit d'une part de proposer des méthodes pour rendre interopérable les solutions proposées dans les trois premiers axes et d'autre part de développer des démonstrateurs.

Il est proposé ainsi de développer une ligne de produits logiciels (FabLab de données) – FabLab virtuel. Cet axe constituera une vitrine scientifique et technologique des actions de recherche déployées par le laboratoire CATAI.

4.1. PARTENAIRES ENVISAGÉS DU MONDE SOCIO-ÉCONOMIQUE :

Les partenaires envisagés du monde socio-économiques sont nombreux. Sans être exhaustifs, on peut nommer les partenaires suivants. Certains d'entre eux ont déjà été approchés dans le cadre du laboratoire CATAI.

- Villes : Bogota, Bucaramanga, Grenoble, Nice, Chalon sur Saône
- Industriels Ville : Bouygues, Vinci, Saint Gobain, Poma, Astom
- Industriels infrastructures : NVidia, Intel, Atos-Bull, Orange, Renault

CONCLUSIONS

L'informatique est au cœur de la transformation urbaine, et le projet CATAI démontre comment des outils tels que la modélisation multi-échelle, la visualisation interactive et le calcul haute performance peuvent améliorer la planification et la durabilité des villes.

D'une autre côte, la coopération franco-colombienne offre un avantage comparatif, combinant contextes urbains variés et compétences complémentaires pour générer des solutions innovantes applicables à différents territoires.

Une conséquence des activités de recherche et développement proposées, est l'identification du citoyen comme l'acteur principal.

Le citoyen devient un acteur central dans les villes intelligentes, via la collecte participative de données, l'interaction avec les systèmes numériques et la codécision, favorisant un urbanisme inclusif.
En termes de technologie, les infrastructures et systèmes informatiques hétérogènes sont essentiels pour traiter efficacement les données massives issues des villes, nécessitant une orchestration fine entre ressources, formats et usages. C'est pour ça que le calcul des hautes performances est relevant.

CATAI agit aussi comme un levier de formation et de transfert, avec des programmes de cocréation et de diffusion de connaissances , renforçant ainsi les capacités locales et la diffusion des innovations.

REMERCIEMENTS

Nous tenons à exprimer notre profonde gratitude à l'Ambassade de France en Colombie, et en particulier à Monsieur Régis Guillaume, attaché académique du service diplomatique français en Colombie, pour son accompagnement précieux et son appui constant à notre initiative. Son engagement a grandement contribué à renforcer la coopération scientifique et culturelle entre nos pays.

Nos sincères remerciements vont également à l'Alliance Française de Bucaramanga, et tout spécialement à Madame Amparo Caballero, dont la disponibilité, la sens
ibilité et l'appui indéfectible ont été essentiels pour la concrétisation de cette démarche. Grâce à le soutien des différents institutions et perssones, ce projet trouve une portée plus humaine, inclusive et durable.
Voici une bibliographie synthétique et sélectionnée (15 entrées maximum) inspirée des thématiques du projet CATAI. Elle combine **sources académiques** et
références web utiles pour approfondir les domaines abordés :

RÉFÉRNCES BIBLIOGRAPHIQUES

1. Batty, M. (2013). *The New Science of Cities*. MIT Press.
2. Croitoru, A., & Nayak, S. (2014). *Geospatial Data and Smart Cities*. In: Smart Cities and Smart Spaces. Springer. Mitchell, W.J. (1999). *e-Topia: Urban Life, Jim—but Not As We Know It*. MIT Press.

3. Townsend, A. (2013). *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. W. W. Norton & Company.
4. Nunes, M., & Camargo, C. (2017). *Urban Computing and Smart Cities*. In Handbook of Smart Cities. Springer.
5. Allwinkle, S., & Cruickshank, P. (2011). *Creating Smart-er Cities: An Overview*.
6. *Journal of Urban Technology*, 18(2), 1–16. Ratti, C., & Claudel, M. (2016). *The City of Tomorrow: Sensors, Networks, Hackers, and the Future of Urban Life*. Yale University Press.
7. Mora, L., Deakin, M., & Reid, A. (2019). *Smart City Development Paths: Insights from 25 Global Cases*. *Cities*, 92, 222–232.

RÉFÉRENCES WEB ET PROJETS LIÉS

1. [MIT City Science Lab](https://cities.media.mit.edu/)
2. [Smart City Institute – Université de Liège](http://labos.ulg.ac.be/smart-city/homepage/)
3. [Smart Cities Center–Columbia University](https://datascience.columbia.edu/smart-cities)
4. [CASA – University College London](https://www.ucl.ac.uk/bartlett/casa)
5. [CityLab@Inria](https://citylab.inria.fr/)
6. [Programme européen ESPON](https://www.espon.eu/) – projets de données spatiales européennes
7. [Open Geospatial Consortium (OGC)](https://www.ogc.org/) – Normes pour la géovisualisation et l'interopérabilité
8. [MIT City Science Lab](https://cities.media.mit.edu/)
9. [Smart City Institute – Université de Liège](http://labos.ulg.ac.be/smart- city/homepage/)
10. 3. [Smart Cities Center–Columbia University](https://datascience.columbia.edu/smart-cities)
11. [CASA – University College London](https://www.ucl.ac.uk/bartlett/casa)
12. [CityLab@Inria](https://citylab.inria.fr/)
13. [Programme européen ESPON](https://www.espon.eu/) – projets de données spatiales européennes
14. [Open Geospatial Consortium (OGC)](https://www.ogc.org/) – Normes pour la géovisualisation et l'interopérabilité

L'INFORMATIQUE POUR LA VILLE INTELLIGENTE ET LE DEVELOPPEMENT DURABLE

JOSE TIBERIO HERNANDEZ PEÑALOSA,1 CARLOS JAIME BARRIOS HERNÁNDEZ2, MICHEL RIVEILL 3, YVES DENNEULIN4, CLAUDIA RONCANCIO4, FRÉDÉRIC MERIENNE5, HAROLD CASTRO1, ISMAEL PEÑA6, HELGA DUARTE6, OSCAR ALBERTO CARRILLO ROZO7, FRÉDÉRIC LE MOUËL8, ARTURO PLATA GOMEZ2, JORGE LUIS CHACÓN2, PABLO FIGUEROA1, RAUL RAMOS POLLÁN2 , GABRIEL RODRIGO PEDRAZA FERREIRA2 ET MIREILLE BIAY-FIALLO3.

1. UNIVERSIDAD DE LOS ANDES (UNIANDES), COLOMBIA
2. UNIVERSIDAD INDUSTRIAL DE SANTANDER (UIS), COLOMBIA
3. UNIVERSITÉ NICE-SOPHIA ANTIPOLIS (UNSA), FRANCE
4. UNIVERSITÉ DE GRENOBLE (UDG), FRANCE
5. ÉCOLE NATIONALE DES ARTS ET MEIIÈRES– PARIS TECH (ENSAM- PARIS TECH), FRANCE
6. UNIVERSIDAD NACIONAL DE COLOMBIA (UNAL), COLOMBIA
7. ÉCOLE SUPÉRIEURE DE CHIMIE, PHYSIQUE, ÉLECTRONIQUE DE LYON (CPE-LYON), FRANCE
8. INSTITUT NATIONAL DES SCIENCES APPLIQUÉES DE LYON (INSA-LYON), FRANCE

RÉSUMÉ

La collaboration franco-colombienne en technologies de l'information avancées, à partir la rechercher et le développement collaboratif, promit le déploiement et l'usage responsable du numérique pour bâtir des villes intelligentes et durables. Elle vise à intégrer des solutions technologiques pour améliorer la gestion urbaine, tout en respectant les ressources naturelles et les principes d'inclusion sociale. Cette approche encourage l'échange de données, la participation citoyenne et l'innovation ouverte. Elle cherche à réduire les inégalités numériques et à renforcer la résilience des territoires face aux défis climatiques et sociaux. Le numérique devient ainsi un levier stratégique pour un développement urbain éthique et collaboratif. Cet article montre les principaux enjeux vis-à-vis la formalisation de la collaboration pour des problématiques de ville durable.

1. INTRODUCTION

La mondialisation des échanges et l'organisation du travail que ce dernier induit, provoquent un flux migratoire important de la population rurale vers les villes. Ce mouvement mondial est encore plus critique dans les villes en développement rapide. La conséquence de cette situation est une énorme difficulté pour gérer de manière pertinente et cohérente la croissance des villes. Des problèmes de natures diverses apparaissent : ceux-ci peuvent concerner directement chacun des citadins mais aussi en amont ceux qui sont en charge de la gestion de la cité. Nous pouvons citer par exemple, les problèmes liés à l'inadéquation des voies de circulation aux flux de voitures, l'augmentation de la pollution ou l'accès des citoyens aux services de la cité.

L'ensemble des partenaires du LIA sont engagés d'une manière ou d'une autre dans ce vaste chantier ou l'informatique (sous ses multiples facettes) peut permettre l'anticipation des problèmes en vue d'apporter des solutions à certains des problèmes de croissance rapide des villes ou d'offrir aux citoyens la possibilité d'être acteur du développement de leur cité. Ce projet de laboratoire international associé franco-colombien CATAI propose de fédérer les activités de recherche et de rendre plus cohérents les divers travaux menés par les uns et les autres, autour de l'apport des travaux menés par les uns et les autres sur l'apport de l'informatique à la ville intelligente avec pour cible de générer de la durabilité urbaine.

La force d'une telle collaboration est d'analyser des situations urbaines différentes avec leurs dynamiques propres en mettant en commun des concepts et outils numériques qui rendent réalistes le développement durable de nos villes.

2. CONTEXTE

Les difficultés induites par la croissance de la plupart des grandes zones urbaines sont fortement dépendantes du continent d'appartenance. En revanche, les confrontations entre villes de natures différentes de leurs problèmes et leurs démarches permettent de proposer des solutions originales.

Des jumelages et programmes communes entre des villes françaises et colombiennes existent (par exemple entre Grenoble et Bucaramanga) permettant d'échanger sur des problématiques communes. Parmi ces problématiques communes, la gestion et l'accès aux données de la ville sont critiques car elles impliquent l'ensemble des acteurs de la ville (autorités, construction, maintenance, gestion, citadin). L'informatique est au cœur de ces problématiques.

Par ailleurs, il existe de nombreuses relations antérieures et actuelles dans le domaine de l'informatique entre la France et la Colombie. Ces relations sont nourries par le fait que de nombreux enseignants-chercheurs Colombiens ont effectué leur thèse en France. Des travaux communs ont déjà eu lieu sur la thématique de l'informatique pour la ville intelligente.

Côté français, 3 laboratoires sont impliqués : I3S, LIG et Le2i qui sont tous 3 des UMRs du CNRS dépendant de l'INS2i. Par ailleurs, aux vues de la thématique abordée, d'autres équipes de recherche pourront être associées aux travaux comme par exemple l'UMR ESPACE (CNRS / UNS) dépendant de l'INEE.

Les raisons évoquées motivent la proposition de la création d'un laboratoire international associé franco-colombien sur cette thématique. La constitution de ce laboratoire permettra de construire dans la durée un programme de recherche pour adresser certains verrous existants, apporter des solutions aux problèmes posés et diffuser ces solutions dans le milieu socio-économique au service du bien-être du citadin.

3. CONSORTIUM

3.1 PRÉSENTATION DE L'ÉQUIPE CATAI

L'équipe CATAI constituée par certains membres des laboratoires présentés ci-après comprendra des compétences complémentaires permettant d'adresser les problématiques de l'informatique au service de la ville intelligente. Par ailleurs, les plateformes technologiques des laboratoires partenaires seront mises au service du laboratoire CATAI. Les moyens suivants sont identifiés :

Domaine	Équipements
Calculateurs hautes performances	Infrastructures de Calcul de Grande Échelle Supercalculateurs (Grappes) Grilles de Calcul Infrastructures de Calcul de Haute Performance Embarquées
Visualisation interactive	Larges écrans de visualisation Salles d'immersion virtuelle Simulateurs de conduite
Centres de stockage et préservation des données	Infrastructures de stockage et de préservation des données spécialisés pour le projet. Serveurs dédiées de webservices et gateways spécialisés

3.2 PARTENAIRES

Les partenaires academiques et scientifiques impliquées dans la proposition, son les suiantes :

3.2.1 LABORATOIRE LE2I (FRANCE)

Le Laboratoire d'Electronique, d'Informatique et de l'Image (Le2i) est une Unité Mixte de Recherche (U.M.R. 6306) rattachée principalement à l'INSII et à l'INSIS en institut secondaire. Le laboratoire comprend 96 Enseignants-Chercheurs, 17 ingénieurs, techniciens et personnels d'encadrement administratif, et environ 95 doctorants et post-doctorants, le Le2i est le seul laboratoire en Sciences et Technologies de l'Information et des Communications (STIC) de la région Bourgogne. Il couvre un large spectre d'activités au sein de trois départements scientifiques : A) Informatique,
B) Electronique et C) Vision. Le Le2i est réparti sur 4 sites en Bourgogne (Dijon, Le Creusot, Auxerre et Chalon-sur-Saône).
Les Etablissements tutelles du Le2i sont :

- Université de Bourgogne
- Arts et Métiers ParisTech (ENSAM)
- CNRS

Le laboratoire Le2i développe des activités de recherche sur la thématique de la ville depuis de nombreuses années à travers l'équipe « réalité virtuelle » et le projet transversal « CheckSem ». L'équipe « réalité virtuelle » à l'institut image a été impliquée dans des projets de recherche mettant en œuvre l'immersion virtuelle au service du patrimoine bâti et du bâtiment (numérisation 3D, maquette virtuelle interactive, immersion virtuelle et réalité augmentée). Deux start-ups ont été essaimées en proximité avec ce champ d'application : la société On-Situ (créée en 2006) et la société Paztec (créée en 2013). Le projet CheckSem s'intéresse à la modélisation sémantique et formelle de connaissances et leur manipulation pour le domaine du BiM sémantique. Une plateforme d'intelligence sémantique a été développée ainsi qu'une architecture orientée services autour d'un noyau combinant des triplestores, du graph mining, des métaheuristiques, et du model checking. Les travaux de recherche de Checksem ont permis en 2008 la création de la société

Active3D.

3.2.2 LABORATOIRE I3S (FRANCE)

Le laboratoire I3S est une unité mixte de recherche Université Nice Sophia Antipolis
/ CNRS associé à Inria. Il rassemble plus de 300 personnes dont 120 chercheurs sénior (enseignants-chercheurs de l'UNS ou chercheurs CNRS (18) et chercheurs Inria (12)) et 80 doctorants.

Le laboratoire I3S est au cœur d'un réseau de partenariats et de programmes collaboratifs, nationaux et internationaux, qui soutiennent et structurent son activité. Ainsi, il est actif dans plusieurs pôles de compétitivité, en particulier : SCS (Solutions Communicantes Sécurisées), PEGASE (Pôle aéronautique et spatial) et PASS (Parfums, Arômes, Senteurs et Saveurs).

Le laboratoire I3S développe depuis plusieurs années des travaux ayant un lien avec la thématique du LIA principalement autour des deux aspects suivants :

- Intelligence artificielle et principalement a) sur les systèmes muti-agent en particulier pour modéliser et simuler des comportements lorsque les agents ont des objectifs (individuels ou non) et des croyances sur le monde dans lequel ils évoluent, ce qui pourrait influencer l'ordre voir même le choix de leurs actions.Ces agents peuvent revoir leurs croyances en fonction de nouvelles informations. Le fait d'utiliser de tels agents dotés d'une représentation de leur état mental permet de modéliser de façon plus fidèle à la réalité les systèmes complexes où les acteurs sont des êtres humains. b) à la fouille de données pour proposer de nouvelles solutions par l'analyse des données passées sur l'évolution de villes
- Génie logiciel et principalement sur les techniques de développement orientées services à base de composants permettant d'augmenter la productivité et la réutilisabilité en découplant les différentes facettes de logiciels. L'ingénierie dirigée par les modèles permet de capitaliser le savoir-L'ingénierie dirigée par les modèles permet de capitaliser le savoir- faire dans les modèles loin des plate-formes technologiques tout en maintenant la traçabilité avec des applications en cours d'exécution. Les problèmes relatifs au passage à l'échelle en terme de stockage et d'accès aux données, de performances dans les calculs et dans les propagations d'information sont particulièrement critiques dans le cadre des applications ciblées. Ainsi les infrastructures à grande échelle deviennent un support indispensable et transparent pour l'exécution d'applications complexes nécessitant un très grand nombre de ressources distribuées à travers des principes de logiciels divers comme les services, le cloud computing et de grid computing, reposant dans le contexte des villes sur des supports très hétérogènes aux performances diverses.

Sera associé au laboratoire I3S, une équipe de recherche du l'UMR ESPACE (4 enseignants-chercheurs - 1 chercheur CNRS) qui développe des travaux autour des interactions multi-échelles et fonctionnement des systèmes territoriaux. L'approfondissement de la connaissance des systèmes territoriaux, et plus spécialement des interrelations espace-environnement-société, constitue l'une des finalités du projet de recherche.

3.2.3 LABORATOIRE LIG (FRANCE)

Le Laboratoire d'Informatique de Grenoble rassemble près de 500 chercheurs, enseignants-chercheurs, doctorants et personnels en support à la recherche. Ils relèvent des différents organismes et sont répartis sur les deux sites du LIG : le campus de Grenoble et Montbonnot.

Le projet scientifique du LIG est l'"Informatique ambiante et durable". L'ambition est de s'appuyer sur la complémentarité et la qualité reconnue des 22 équipes de recherche du LIG pour contribuer au développement des aspects fondamentaux de la discipline (modèles, langages, méthodes, algorithmes) et pour développer une synergie entre les défis conceptuels, technologiques et sociétaux associés à cette thématique que l'on retrouve très présente dans le projet CATAI.

Le LIG contribuera à CATAI sur les aspects gestion de données à large échelle, calcul intensif et

mouvements de données associés ainsi que sur le traitement des informations spatio-temporelles, depuis leur acquisition au moyen de capteurs physiques ou citoyens ou par le biais de sources de données officielles, ou bien encore liées et ouvertes, jusqu'à leur restitution sous diverses formes : cartes, graphes et graphiques dynamiques et interactifs, jeux de données structurés ou non structurés, services web géographiques, etc., sans omettre l'association de métadonnées renseignant sur la provenance de ces données et sur leur qualité intrinsèque au regard d'un certain nombre de critères définis.

3.2.4 LABORATOIRE IMAGINE/COMIT-UNI ANDES (COLOMBIE)

Les équipes de recherche IMAGINE (Informatique visuelle) et COMIT (Communications et Technologies de l´information), sont rattachées au département d´Ingénierie Informatique (Sistemas y Computación) de l´Université de los Andes, â Bogotá. L´équipe IMAGINE comprend 4 Enseignants-Chercheurs, et 7 doctorants et post-doctorants. L´équipe COMIT comprend 9 Enseignants-Chercheurs, et 12 doctorants et post-doctorants.

Le laboratoire Imagine/Comit développe des activités de recherche sur la thématique de la ville depuis de nombreuses années â travers des projets de visual analytics et bigdata analysis Des collaborations avec des équipes de génie de transport, et d´urbanisme ont données lieu â des projets appliqués â la ville de Bogota.

Cette dynamique a permis la création de nouvelles sociétés, particulièrement la société Datatraffic (créée en 2007) qui offre des services de valeur ajouté supportés par une cartographie digitale augmentée Le projet TaCAT permet la mise en œuvre de nouveaux outils interactifs pour la prise des décisions dans le domaine des systèmes urbains. Des particularités des sources d'information, quantité et qualité des données, différents outils d'analyse et différents stakeholders pour la prise de décisions, font partie des caractéristiques de ce projets.

IMAGINE/COMIT va contribuer à CATAÏ dans les domaines de bigdata, Visual Analytics, Internet of things, Realité virtuelle et IHM notamment. L'analyse des données spatio temporelles, les nouveaux environnements interactifs d'analyse collaboratif, l´integration des données heterogenes et des outils d'analyse HPC nous offrent des opportunités exceptionnelles de collaboration scientifique.

3.2.4 CENTRE DE CALCUL D'HAUT PERFORMANCE ET SCIENTIFIQUE DE L'UNIVERSITÉ INDUSTRIELLE DE SANTANDER (SC3UIS) (COLOMBIE)

Le Centre de Calcul d'Haut Performance et Scientifique a été crée en 2011 à l'Université Industrielle de Santander dans la cadre du projet « Technologies avancées de l'information et la communication pour la science et la génie de l'orient colombien », Le processus de création du centre à compromis un accompagnement de côte français vis à vis le Laboratoire d'Informatique de Grenoble, l'INRIA Rhône Alpes y le Laboratoire d'Informatique Signaux et Systèmes de Sophia Antipolis.

La collaboration des laboratoires et institutions français à permis la définition des lignes de travaux de recherche et développement, une planning d'évolution du centre à cinq ans de façon opérative y administratif et la planification des lignes de formation articulés avec les programmes d'école d'ingénieur, master, doctorat en informatique et formation continue.
Le centre SC3UIS a comme objectifs le support de projets de recherche, développement et innovation dans

les moyennes academique et industrielle. Le support involucre des activites de formation dans different niveaux (utilisateurs scientifiques, developpeurs, administrateurs d'infrastructure, etc) et aussi la médiation strategie entre entreprises, les institutions de gouvernement et le secteur scientifique. Son activité il n'est pas uniquement addressé aux problematiques regionaux, sinon colombiennes et en projets internationaux, principalement en amerique latine.

Le centre SC3UIS au niveau scientifique a deux importants composants: un qui garanti une activitié autour des sciences de la communication et l'information avec des lignes de recherche en architectures de grande echelle, analytique des données

- donc le big data, l'architecture et génié de logiciel, le calcul de haute performance, la programmation concurrente et parallele, des systemes distribués et les systèmes qui supportent le processus parallele. L'autre s'occupe des applications scientifiques et la recherche au tour des problèmes en science et ingenierie qu'on besoin de calcul de haute performance: l'eau et l'environnement, l'astronomie et l'astrophysique, les sciences de la vie, les sciences sociaux et humaines, l'energie, l'eau et le gaz, le traitement des données physique-chimie et les sciences de la terre.

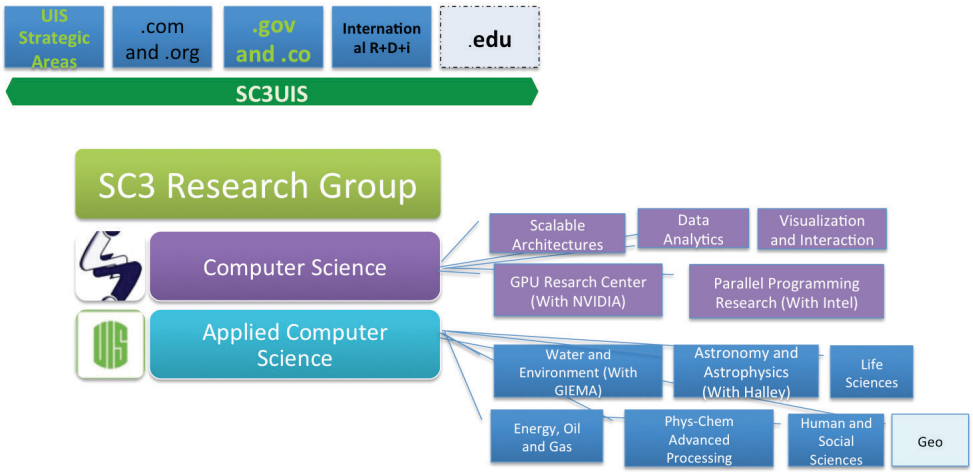


Figure 1. organisation de l'activité scientifique a supprime : X

3.2.6 GRUPO DE INVESTIGACIÓN GEOURBE DE LA UNIVERSIDAD NACIONAL DE COLOMBIA (COLOMBIE)

Le groupe de recherche GEOURBE regroupe des chercheurs intéréssés par les thématiques urbaines et régionales, à partir d'une approche géopolitique basée sur l'idée selon laquelle la ville, et par extension toute unité territoriale, est un produit social résultant d'intérêts et de valeurs sociales conflictuelles.

Le groupe se compose essentiellement de géographes, mais il est ouvert à d'autres disciplines reliées à la géographie par l'intérêt dans la dimension spatiale des processus sociaux. Il est inscrit dans la Red Internacional de Fuentes de Información y Conocimiento para Gestión de Ciencia, Tecnología e Innovación (SCienTI)de COLCIENCIAS.

3.2.7 LABORATOIRE “ETUDE DES STRUCTURES, DES PROCESSUS D’ADAPTATION ET DES CHANGEMENTS DE L’ESPACE” (ESPACE) - UMR 7300 CNRS

ESPACE associe le CNRS et des laboratoires de trois universités : Nice Sophia-Antipolis, Aix Marseille Université et Avignon. Antipolis, Aix Marseille Université et Avignon. Il est rattaché à deux instituts du CNRS, l’Institut National des Sciences Humaines et Sociales et l’Institut National Ecologie et Environnement.

L’UMR ESPACE dirige l’Observatoire Homme-Milieu Littoral méditerranéen et appartient au Laboratoire d’Excellence « Dispositif de Recherche Interdisciplinaire pour les Interactions Hommes-Milieux » (DRIIHM). ESPACE-Nice rassemble 16 enseignants-chercheurs et chercheurs CNRS, 9 ingénieurs et techniciens et 17 doctorants.

Les travaux du laboratoire ESPACE sont principalement consacrés aux espaces urbanisés plus ou moins denses, à l'analyse de leur organisation, de leur évolution, de leurs interactions avec les systèmes naturels et agricoles avec lesquels ils sont en contact. Ces espaces en mutation constante sont étudiés selon une approche systémique qui permet d'appréhender de manière globale les divers champs d'étude, mais aussi d'individualiser les éléments et les facteurs qui semblent pertinents pour étudier cette complexité.

Les phénomènes urbains sont analysés dans leur dimension multi-niveaux afin de déceler les types de relations inter-scalaires et déterminer le rôle des jeux d'échelles dans le fonctionnement des systèmes territoriaux. Les travaux portent sur la caractérisation des géosystèmes urbains et leur trajectoire déclinée selon deux concepts : le déploiement (conservation de l'équilibre) et l'évolution (transition d'un état à un autre).

Utilisant les outils d’analyse spatiale et les SIG, le laboratoire s’est très tôt intéressé à la modélisation t à la simulation dans un but heuristique : il fut notamment leader sur les approches individus-centrées en géographie dès 1998 avec un programme de recherche où des modèles de simulation de la mobilité résidentielle ont été réalisés à partir d’approches utilisant les automates cellulaires et les systèmes multi-agents. agents. La géoprospective, c’est-à-dire l’anticipation des changements spatiaux par différents types de modélisation et de simulation, est une des spécialités du laboratoire.

4. PROJET SCIENTIFIQUE

4.1 CONTEXTE ET ENJEUX

Le projet scientifique proposé consiste à faire collaborer des unités de recherche ayant des compétences complémentaires dans divers domaines de l’informatique au service de la ville intelligente et la soustenibilité, centrées dans les besoins humains. La ville de demain aura des besoins cruciaux dans la constitution, la gestion et l’accès à ses données. Ainsi, la ville doit avoir son modèle numérique évoluant avec elle.

Les données collectées peuvent être apportées par des sources diverses et très hétérogènes au niveau de leurs fiabilités. La participation active des citadins permettra de mettre le citoyen au cœur de la ville tant pour informer que pour s’informer et co-décider. Les services induits par cette ville inter-connectée pourront être optimisés (service à la personne, sécurité, mobilité, accès aux infrastructures, guide touristique…). Les co-décideurs de leur environnement urbain auront accès à des systèmes de visualisation interactive

pour l'état des lieux et l'état projeté (simulation) des mobilités (temps de trajets), pollution, économie des quartiers

Les enjeux de ces besoins sont au carrefour de différents métiers de l'informatique : la constitution de base de données avec traçabilité, la gestion de masses de données lourdes, hétérogènes et multi-échelles, la visualisation interactive de données complexe.

4.2 POSITIONNEMENT

Sur la thématique de la ville du futur, plusieurs universités prestigieuses développent des programmes de recherche ; preuve que les enjeux sociétaux de la ville sont prégnants dans le monde. Le tableau suivant synthétise quelques laboratoires (tous relativement jeunes) existants dans le domaine. La liste n'est bien sûr pas exhaustive.

laboratoire	Objectifs
City Science Initiative MiT Media Lab (USA) http://cities.media.mit.edu/	To build the cities that the world needs, we need a scientific understanding of cities that considers our built environments and the people who inhabit them. Six initial themes represent a cross section of the interdisciplinary research that will be undertaken to address the major challenges associated with global urbanization. Urban analytics and modeling Incentives and governance Mobility networks Places of living and work Electronics and social networks Energy networks
Smart City Lab University of Bologna (Italy) http://smartcity.csr.unibo.it/	Research in urban ict innovation for a better life. The "Smart City" model is strictly related to the ICT infrastructures that the city itself shares with the citizens; these innovative architectures provide a more efficient conservation of the fixed heritage and a better standard of living, thanks to modern and full-accessible services.
Smart City Institute Accenture, Belfius, Ville de Liège, Université de Lière (Belgique) http://labos.ulg.ac.be/smart-city/homepage/	The researchers of the Smart City Institute work on the development of a scientific expertise – with an international perspective and impact – on the management of smart cities.
City Lab Inria (France) https://citylab.inria.fr/	The Inria Lab CityLab@Inria studies ICT solutions toward smart cities that promote both social and environmental sustainability. A strong emphasis of the Lab is on the undertaking of a multi-disciplinary research program through the integration of relevant scientific and technology studies, from sensing up to analytics and advanced applications, so as to actually enact the foreseen smart city Systems of Systems.

Smart cities research Center Univ of California Berkeley (USA) http://smartcities.berkeley.edu/	The Research Center was established at UC Berkeley as a part of inter-disciplinary effort to advance quantitative modelling of urban systems. It's core is at CEE Systems and Transportation. We work on mathematical models and data analytics with public agencies as well as private companies and focusing on fundamental research in a broad area of Smarter Cities.
Smart Cities Center Columbia University (USA) http://datascience.columbia.edu/smart-cities	Research conducted by the Smart Cities Center develops and monitors sustainable urban infrastructure and buildings, improves the power supply through smart grid technology, detects and counteracts problems with aging urban infrastructure, calculates and communicates optimal transportation routes under congested traffic conditions, and deploys ubiquitous sensing devices to facilitate everyday activities in a crowded urban environment.
Laboratoire	Objectifs
City Science Initiative MiT Media Lab (USA) http://cities.media.mit.edu/	To build the cities that the world needs, we need a scientific understanding of cities that considers our built environments and the people who inhabit them. Six initial themes represent a cross section of the interdisciplinary research that will be undertaken to address the major challenges associated with global urbanization. Urban analytics and modeling Incentives and governance Mobility networks Places of living and work Electronics and social networks Energy networks
Smart City Lab University of Bologna (Italy) http://smartcity.csr.unibo.it/	Research in urban ict innovation for a better life. The "Smart City" model is strictly related to the ICT infrastructures that the city itself shares with the citizens; these innovative architectures provide a more efficient conservation of the fixed heritage and a better standard of living, thanks to modern and full- accessible services.
Smart City Institute Accenture, Belfius, Ville de Liège, Université de Lière (Belgique) http://labos.ulg.ac.be/smart-city/homepage/	The researchers of the Smart City Institute work on the development of a scientific expertise – with an international perspective and impact – on the management of smart cities.
City Lab Inria (France) https://citylab.inria.fr/	The Inria Lab CityLab@Inria studies ICT solutions toward smart cities that promote both social and environmental sustainability. A strong emphasis of the Lab is on the undertaking of a multi- disciplinary research program through the integration of relevant scientific and technology studies, from sensing up to analytics and advanced applications, so as to actually enact the foreseen smart city Systems of Systems.

Smart cities research Center Univ of California Berkeley (USA) http://smartcities.berkeley.edu/	The Research Center was established at UC Berkeley as a part of inter-disciplinary effort to advance quantitative modelling of urban systems. It's core is at CEE Systems and Transportation. We work on mathematical models and data analytics with public agencies as well as private companies and focusing on fundamental research in a broad area of Smarter Cities.
Smart Cities Center Columbia University (USA) http://datascience.columbia.edu/smart-cities	Research conducted by the Smart Cities Center develops and monitors sustainable urban infrastructure and buildings, improves the power supply through smart grid technology, detects and counteracts problems with aging urban infrastructure, calculates and communicates optimal transportation routes under congested traffic conditions, and deploys ubiquitous sensing devices to facilitate everyday activities in a crowded urban environment.
CASA (The Centre for Advanced Spatial Analysis) UK. http://www.bartlett.ucl.ac.uk/casa	CASA is engaged in generating new knowledge and insights for use in city planning, policy and design and drawing on the latest geospatial methods and ideas in computer-based visualisation and modelling.

4.3 STRATÉGIE SCIENTIFIQUE

La stratégie scientifique du laboratoire CATAÏ aura pour contexte les aspects liés à l'évolution rapide des villes soit pour offrir des outils de planification ou de simulation aux décideurs, soit pour offrir des outils participatifs aux citoyens. En effet, les outils informatiques peuvent s'avérer particulièrement utile dans quatre thématiques : le transport, la santé publique, les réseaux de communication, la gestion des risques.

L'originalité de la proposition réside dans la complémentarité des partenaires pour adresser cette problématique tout en ayant un socle commun de compétences. Compte tenu des enjeux identifiés, le laboratoire CATAI focalisera son action sur la chaîne de la valeur depuis la constitution des données de la ville, leur gestion jusqu'à leur visualisation interactive. Compte tenu du caractère applicatif de la proposition, des compétences en génie logiciel seront mis en œuvre de façon à proposer des démonstrateurs technologiques.

Ainsi, les compétences des différents partenaires seront mutualisées pour adresser 4 champs permettant de couvrir la chaîne de valeur identifiée :

- Visualisation interactive
- Modèle multi-échelle
- Calcul de haute performance
- Génie Logiciel

Par ailleurs, les compétences des partenaires impliqués dans le laboratoire CATAI ne seront pas suffisantes pour embrasser l'ensemble des problématiques scientifiques envisagées qui induisent une approche fortement pluri-disciplinaire.

Aussi, les travaux seront menés en partenariat avec des laboratoires spécialisés dans l'étude de ces domaines applicatifs (géographie, urbanisme, architecture, histoire, sociologie, ergonomie, psychologie cognitive…).

4.4 AXES THÉMATIQUES

Les thèmes identifiés correspondent aux 4 axes thématiques résumés dans la figure suivante. Les dimensions multi-modèles et génie logiciel sont transverses à l'ensemble du projet et donc les interfaces entre le calcul haute performances et la visualisation.

Axe 1 – Modélisation multi-échelle

L'objectif de cet axe de recherche est d'aboutir à des modélisations de la ville permettant de maîtriser la complexité de l'existant et de son évolution et de proposer des services innovants et des connaissances utiles à la prise de décision en matière de politiques urbaines. Deux thèmes font l'objet des modélisations dans cet axe : la modélisation des données et des connaissances et la modélisation spatio-temporelle des formes et des processus urbains.

Sous-axe 1. Modélisation de données et des connaissances sur la ville

D'un point de vue de la gestion des données, l'enjeu scientifique est de définir des processus de migration de sources de données traditionnelles (systèmes d'information classiques, données structurées) pour qu'elles eviennent exploitables dans une logique ouverte au service de la ville et des citoyens.

L'objectif est de pouvoir tirer profit de telles sources conjointement avec des données actuelles produites et stockées avec des technologies récentes (systèmes NoSQL, approches à base de connaissance). Il y aura à maitriser une masse de données combinant données structurées, non structurées, textuelles et une dimension spatio-temporelle presque omniprésente. Ces caractéristiques, qu'on qualifierait de big data, sont accentuées d'autant plus par la dimension ubiquitaire des systèmes déployés au service des villes intelligentes.

Les recherches concernent la maîtrise de la collecte et l'exploitation de données relevées par des capteurs physiques ou logiciel, couplés à des stratégies où l'on contrôle fortement la production des données ou au contraire des approches de type crowdsourcing basée sur une participation très subjective, et quasi opportuniste des habitants.

Diverses échelles de collecte et de traitement devront être possibles afin de répondre aux besoins et aux contraintes de coût et qualité des utilisateurs. Les citoyens-capteurs ou produsers, à l'image des processus de démocratie participative, pourront être sollicités pour des campagnes de collectes de données ou encore des enquêtes d'opinion.

Ces sondages nécessitent des infrastructures de données capables : i) en amont, de sélectionner au besoin sur critères divers – degré d'expertise, localisation, âge, profession, etc. – tout ou partie d'une population, de qualifier les données ainsi collectées, de construire ou compléter les métadonnées indispensables, de lier ces données et métadonnées avec des données institutionnelles, éventuellement de les assembler en des séries temporelles longues et surtout cohérentes ; ii) en aval de les restituer à la fois à des experts pour que soient menées des analyses conduisant à la prise de décision dans les domaines de l'urbanisme, de l'aménagement du territoire, de la prévention des risque, de la santé etc., mais également aux citoyens, partie prenante dans ce processus de production et d'exploitation des données de la ville, à des fins d'information.

Si les technologies actuelles, notamment celles préconisées par l'Open Geospatial Consortium, permettent, en principe, la construction de telles infrastructures, la mise en place d'un tel flux de données combinant données institutionnelles et citoyennes, reste largement un défi à relever.

D'un point de vue des systèmes d'information, comprendre et modéliser des socio- écosystèmes complexes reste un problème non maitrisé. Les systèmes « supportant » les villes intelligentes sont intrinsèquement complexes dans le sens où ils sont très dynamiques et combinent un éventail très large de services et d'acteurs appartenant à des organisations hétérogènes et autonomes. Les approches de conception de systèmes socio-techniques complexes constituent également un sujet de recherche à explorer.

Sous-axe 2. Modélisation des formes et des processus urbains de changement rapide

Les transformations rapides des espaces urbanisés dans le monde exigent une planification capable de faire face aux défis économiques, sociaux et environnementaux de la ville d'aujourd'hui. Cette planification doit être adaptée à la complexité des phénomènes urbains, mais aussi au rythme des évolutions qui s'imbriquent dans différentes échelles spatiales (des quartiers aux réseaux de villes) et temporelles (de la mobilité quotidienne aux transitions urbaines et métropolitaines). Les recherches actuelles en modélisation urbaine tendent à se focaliser sur le phénomène d'extension des aires urbaines et sur les conséquences qui en résultent, notamment sur le plan environnemental. Or, dans nombre de pays du Sud, et dans une moindre mesure, du Nord, le processus de croissance urbaine se manifeste également par des transformations de la structure urbaine qui touchent les centres comme les périphéries déjà urbanisées. Ce réaménagement continuel ne relève pas uniquement d'opérations d'urbanisme. Des transformations spontanées, en lien ou non avec les actions programmées, pouvant se produire à des rythmes très rapides, transforment la ville dans ses dimensions horizontales et verticales comme dans ses fonctions. Dès lors, l'anticipation de la régénération intra-urbaine et de l'étalement urbain, par la modélisation et la simulation, est d'un intérêt majeur.

La modélisation urbaine actuelle se limite souvent, soit à la simulation des mobilités quotidiennes en fonction de la répartition des localisations des populations et des fonctions urbaines dans des modèles de type LUTI (Land Use/Transport Interaction Models), soit à la simulation de l'évolution de l'occupation du sol par télédétection et automates cellulaires pour étudier l'extension de la tache urbaine. Or, il est rare de concevoir des modèles qui offrent la possibilité de faire le lien entre les différentes échelles spatiales et temporelles et de formaliser le lien entre les formes et le fonctionnement de la ville. La modélisation urbaine se voit donc confrontée à plusieurs difficultés.

Une première difficulté provient de la nécessité de définir les échelles appropriées à la description des différents processus urbains. Les phénomènes de métropolisation doivent être analysés à différentes échelles spatiales, du quartier au réseau de villes. De façon analogue, le rôle des technologies de l'information et de la communication dans la mobilité quotidienne et de la complexité de l'organisation spatiale des territoires urbains impose une imbrication des échelles temporelles dans les analyses.

La deuxième difficulté provient de la relation entre les différentes échelles. La modélisation cherche souvent à décrire l'auto-organisation de la ville par l'interaction entre ses éléments constitutifs dans une approche « bottom-up ». Or, il est indéniable qu'il est indispensable d'intégrer dans cette modélisation le rôle des politiques publiques, des agents économiques et des facteurs externes (modèle de développement économique, caractéristiques physiques et géographiques, histoire urbaine, etc.) sur le comportement du système correspondant à une approche « top- down ».

La troisième difficulté est celle de l'intégration des flux grandissants de données dans la calibration et la validation de modèles de plus en plus complexes. La modélisation informatique se voit aussi confrontée à la nécessité d'apporter des réponses en temps réel permettant d'ajuster les infrastructures urbaines

au comportement changeant de la ville. La disponibilité d'informations a transformé la manière dont les modèles sont conçus et leur rôle dans la prise de décision par les décideurs publics. Ils jouent un rôle important aussi dans la représentation et la communication des phénomènes. Ils se doivent donc d'assurer leur intelligibilité auprès des différents acteurs de la ville, ce qui impose de nouveaux défis dans la visualisation informatique.
Les objectifs des différents types de modélisation de l'évolution urbaine dans le cadre de ce projet sont :

- décrire les trajectoires des dynamiques urbaines en intégrant la croissante complexité spatio-temporelle de leur fonctionnement, notamment dans le contexte des fast changing cities colombiennes
- mettre en relation les formes urbaines et le fonctionnement de la ville en termes de mobilité (quotidienne et résidentielle), organisation socio-spatiale et impact environnemental
- fournir à la population civile et aux décideurs des outils, notamment cartographiques, pour anticiper les changements dans un contexte d'évolution rapide des formes urbaines
- identifier la possible émergence de phénomènes nouveaux à travers l'analyse de signaux faibles voire contradictoires dans l'évolution de la ville
- étudier les transformations des formes urbaines et pas seulement la croissance de la tache urbaine à partir de simulations à base d'agents morphologiques
- analyser les formes urbaines, non seulement à partir de leur extension surfacique, mais aussi dans les relations allotopiques induites par les réseaux de circulation et dans leur composante verticale à travers des modélisations en 3D
- intégrer des formalismes à base d'incertitude pour faire face à des données manquantes ou multi-sources, aux problèmes d'appréhension d'objets et de schémas spatiaux et à la subjectivité inhérente à la représentation des phénomènes urbains
- développer des passerelles (protocoles de formalisation, modélisation des connaissances, intégration aux règles de la géo-simulation, validation des modèles, etc.) entre les approches modélisatrices à la ville et les connaissances expertes des villes colombiennes produites par des approches plus classiques en géographie et en urbanisme.
- Développer des infrastructures de données spatiales et temporelles, capables de stocker et de restituer ces données dans toute leur diversité, en maîtrisant leur imperfection, en améliorant continuellement les processus de production et de collecte, associés, et, de là, d'en garantir une utilisation avertie.

Axe 2 – Calcul de haute performance

Le modèle ainsi constitué sert de base de travail pour réaliser des simulations et de traitement des donnes masives permettant de représenter ce qui ne se voit pas forcément rapidement ou de réaliser un état projeté (simulation d'indicateurs en fonction d'hypothèses d'aménagement urbain). Des simulations avec de nombreux couplages seront réalisées.

La complexité des données et le nombre de variables induiront des verrous scientifiques liés au calcul haute performance sur de grandes quantités de données dont le stockage et l'accès est aussi un enjeu. La nature incertaine de certaines de ces données pourra également faire l'objet de proposition de méthodes de calculs dédiées. L'évolution de la ville pourra ainsi être projetée et permettra d'aider les co-décideurs dans l'aménagement urbain. Des verrous importants sont également à prévoir pour faire le nécessaire dialogue des méthodes et outils propres à chaque métier intervenant dans le changement de la ville (architecte, urbaniste, ingénieur).

Sous-axe 2.1 - : Infrastructures heterogenes de calcul de haute performance et efficience energetique de calcul:

Le support technologique pour l'execution des modeles de simulation et des applications qui permettent l'execution des logiciels pour le traitement des grandes volumes des donnees, posent des differents questions associés avec le performance, l'efficience en la consommation energetique des processus (vis à vis l'exascale), la portabilitié et la possiblité de utiliser des platforms heterogenes pour accelerer les calculs ou les faire de facon embarqué (traitement des données in situ), toujours en garantir le haute performance.

L'utilisation de differents infrastructures suivent des importants variations dans la construction des algorithms, les mecanismes d'implementation des algorithms, les langages de programmation et l'evaluation de performance. Dans une autre cote, plus relié à l'architecture materielle et sa liaison avec le logiciel, il existe des particularités au niveau de compilation, de intergiciel et de systemes d'exploitation.

Sous-axe 2.2 - Architectures de grande echelle:

Les architectures de grande echelle impliquant des composants logiciels et materiels pour le traitement intensif et distribué des donnees, sur des platforms grid ou cloud. Les problematiques associées, comme la gestion de l'information, la securité, l'interoperabilité entres autres, permettent ne seulement l'interaction sinon le passage à l'echelle des problèmatiques et des données.

Le modèle de visibilité de cloud (IaaS, PaaS et Saas) presentent de façon principale l'interaction technologie-humain et la rélation données-implementation, associées aux problèmes technologiques (langages de programmation, observation de performance, coherence des données et de procées, tolerance aux fauts, etc), que sont dérivées de l'utilisation de platforms de grande echelle.

Axe 3 – Visualisation interactive

L'axe 3 s'intéressera plus spécifiquement aux interfaces permettant aux acteurs de la ville un accès aux données. Une maquette virtuelle de la ville et un environnement de cartographie dynamique basé sur les principes de la géovisualiastion, constitueront ainsi des outils au service de la collaboration entre urbaniste, citoyen, ingénieur, architecte, décideur.

Par ailleurs, la conception et mise en œuvre dénvironnements interactifs de géovisualisation pour l'analyse exploratoire et l'aide à la décision jouera un rôle d'intégrateur entre les partenaires dans les actions du LIA. Cet axe de recherche est structuré en 5 sous-axes suivants, mais aussi il est très attaché à l'axe 2, du sur le besoin de visualisation remote pour des ambiances de collaboration, avec la possibilité d'un calcul et traitement des donnés in situ ou viceverse. La grande echelle alors involucre la proposition des services assocités au haut performance, comme les niveaux des services cloud et la visualisation interactive avancée et le travail collaboratif exprimes dans des sections suivantes.

Sous-axe 3.1 - Extraction de maquette virtuelle :

A partir des modèles conçus dans l'axe 1, une maquette virtuelle doit être construite de façon à permettre une visualisation interactive de la ville intégrant les données. Compte tenu du caractère interactif de la visualisation, les données doivent être simplifiées tout en possédant les éléments justes nécessaires à leur utilisation. La problématique de ce sous-axe est donc liée à l'extraction de maquettes virtuelles à partir des modèles multi-échelles et exhaustifs représentant la ville.

Il s'agit par conséquent de développer des méthodes et outils permettant l'adaptation des données à l'application et aux utilisateurs de l'application. Pour mener à bien cette adaptation des données, des

critères seront recherchés en liaison avec le profil de l'utilisateur ou des utilisateurs de la maquette virtuelle créée et avec les caractéristiques de l'application (visite virtuelle, revue de projet…). Par exemple, le profil de l'utilisateur pourra guider des besoins en calculs (par exemple de flux, de distances, d'hypothèses urbanistiques) qui seront par la suite proposés à l'utilisateur dans sa session de visualisation avancée.

Sous-axe 3.2 - Métaphores de visualisation :

Les données nécessaires à la compréhension de la ville sont plurielles et complexes et ne se limitent pas à sa géométrie. Il peut s'agir de résultats de calculs (liés à la géométrie) ou des hypothèses ou modifications urbaines. Il peut également s'agir d'informations sur la qualité des données (en termes d'incertitude ou de qualification du fournisseur de la donnée).

Ce sous-axe s'attachera à proposer des techniques de visualisation et plus particluièrement de géovisualisation permettant de représenter une variété possibles de données et d'indicateurs en fonction du contexte d'utilisation, des objectifs attendus (communication, analyse, exploration) et des utilisateurs finaux (publics, experts …), ainsi que des différents types de modélisation urbaines. Des techniques existent dans la littérature: multifenêtrage synchronisé intégrant des représentations cartographiques, graphiques et/ou multimédia; cartographie dynamique et animée pour la représentation des dynamiques des territoires, cube spatio-temporel pour la représentation des mobilités urbaines…. Toutefois, face à la densité de données à représenter et à leur hétérogénéité, il est nécessaire de les faire évoluer, tant sur le plan des variables visuelles (couleur, forme, transparence, étiquettes de valeur ...) et des variables dynamiques (clignotement, déplacement, apparition, durée …) habituellement utilisées que sur celui des modes d'expression visuelles. Par ailleurs, la combinaison des modes et des techniques de visualisation au sein d'un même environnement n'est pas aisée à mettre en œuvre pour satisfaire aux exigences de l'utilisateur et à la complexité des données.

Sous-axe 3.3 - Techniques de navigation et interaction :

Les données peuvent être très importantes pour ce qui concerne leurs quantités, ce qui suppose la possibilité de naviguer dans l'espace virtuel des données. Des techniques de navigation seront explorées et évaluées dans le contexte de la ville. En fonction du profil de l'utilisateur, de l'application et du système de visualisation, la navigation pourra se faire avec une vue en miniature de la ville (survol de la ville) ou vue en position de piéton dans la ville. Les techniques de navigation pourront être la technique de la carte virtuelle (avatar google map), la technique du monde en miniature, la technique en navigation à la première personne ou toute autre technique existante dans la littérature.

Une technique de navigation mal maîtrisée sur le sujet peut induire le mal de simulateur. Les techniques de navigation sont gérées par le mapping entre les mouvements de l'utilisateur (doigts sur un Joystick par exemple ou mouvement plus naturel) et les profils de vitesse et accélération de la navigation. Les effets de la technique de navigation et du mapping sur la navigation du sujet pourront faire l'objet de travaux de recherche. Lors de l'expérience de visualisation interactive, l'utilisateur sera également amené à interagir avec les données.

Les modalités d'accès aux données pertinentes pour son utilisation en fonction de son profil constituent un verrou important. Les façons d'adresser les requêtes sur les données et d'accéder aux données demandées seront étudiées. De la même façon, lors d'une session de travail collaboratif sur les données, il serait intéressant de proposer aux utilisateurs des moyens d'annoter les données pour conserver une mémoire du travail de collaborations. Les méthodes et outils d'annotation et la remontée des annotations vers les données natives pourront faire l'objet d'études particulières.

Sous-axe 3.4 – Techniques de visualisation interactive avancée :
En fonction de l'application, différents systèmes de visualisation interactive pourront être mis en œuvre. On pense aux écrans immersifs permettant une visualisation de la maquette numérique à grande échelle. L'utilisateur peut être immergé dans la maquette virtuelle de la ville à l'échelle 1 et partager son expérience avec d'autres utilisateurs. Des systèmes collaboratifs plus simples de type table tactile pourront être mis en œuvre pour des applications particulières également. Ces deux dispositifs sont utilisés dans un contexte de type laboratoire ou bureaux d'études. Ils permettent une bonne immersion avec les données grâce à leurs grands écrans. Un autre type de dispositif permettant la visualisation sur le site de la ville sera étudié. Il s'agit de la réalité augmentée qui permet une interaction in situ avec les données numériques. La problématique est de permettre une adéquation entre les données et le site réel visualisé.

Sous-axe 3.5 - Travail collaboratif :
Un intérêt majeur de ces dispositifs complexes de visualisation interactive avancée réside dans la possibilité de mettre plusieurs utilisateurs de profils différents en collaboration avec les données. Les problématiques posées par ce verrou sont relatives à l'accès aux données lors d'une visualisation sur site par la réalité augmentée ou lors de visualisation cartographique, l'échange de données entre dispositifs distants ainsi que l'adaptation des données en fonction du dispositif dans le cas de collaboration distantes entre systèmes asymétriques. Des protocoles de collaborations devront être définis au préalable.

Axe 4 – Génie logiciel
L'axe 4 concerne le génie logiciel. Cet axe est transversal aux 3 axes précédents et sert de colonne vertébrale aux actions du laboratoire CATAI. Il s'agit d'une part de proposer des méthodes pour rendre interopérable les solutions proposées dans les trois premiers axes et d'autre part de développer des démonstrateurs. Il est proposé ainsi de développer une ligne de produits logiciels (FabLab de données) – FabLab virtuel. Cet axe constituera une vitrine scientifique et technologique des actions de recherche déployées par le laboratoire CATAI.

4.1. PARTENAIRES ENVISAGÉS DU MONDE SOCIO-ÉCONOMIQUE :

Les partenaires envisagés du monde socio-économiques sont nombreux. Sans être exhaustifs, on peut nommer les partenaires suivants. Certains d'entre eux ont déjà été approchés dans le cadre du laboratoire CATAI.

- Villes : Bogota, Bucaramanga, Grenoble, Nice, Chalon sur Saône
- Industriels Ville : Bouygues, Vinci, Saint Gobain, Poma, Astom
- Industriels infrastructures : NVidia, Intel, Atos-Bull, Orange, Renault

CONCLUSIONS

L'informatique est au cœur de la transformation urbaine, et le projet CATAI démontre comment des outils tels que la modélisation multi-échelle, la visualisation interactive et le calcul haute performance peuvent améliorer la planification et la durabilité des villes.
D'une autre côte, la coopération franco-colombienne offre un avantage comparatif, combinant contextes urbains variés et compétences complémentaires pour générer des solutions innovantes applicables à différents territoires.
Une conséquence des activités de recherche et développement proposées, est l'identification du citoyen comme l'acteur principal. Le citoyen devient un acteur central dans les villes intelligentes, via la collecte participative de données, l'interaction avec les systèmes numériques et la codécision, favorisant un urbanisme inclusif.

En termes de technologie, les infrastructures et systèmes informatiques hétérogènes sont essentiels pour traiter efficacement les données massives issues des villes, nécessitant une orchestration fine entre ressources, formats et usages. C'est pour ça que le calcul des hautes performances est relevant.
CATAI agit aussi comme un levier de formation et de transfert, avec des programmes de cocréation et de diffusion de connaissances , renforçant ainsi les capacités locales et la diffusion des innovations.

REMERCIEMENTS

Nous tenons à exprimer notre profonde gratitude à l'Ambassade de France en Colombie, et en particulier à Monsieur Régis Guillaume, attaché académique du service diplomatique français en Colombie, pour son accompagnement précieux et son appui constant à notre initiative. Son engagement a grandement contribué à renforcer la coopération scientifique et culturelle entre nos pays.
Nos sincères remerciements vont également à l'Alliance Française de Bucaramanga, et tout spécialement à Madame Amparo Caballero, dont la disponibilité, la sensibilité et l'appui indéfectible ont été essentiels pour la concrétisation de cette démarche. Grâce à le soutien des différents institutions et perssones, ce projet trouve une portée plus humaine, inclusive et durable.
Voici une bibliographie synthétique et sélectionnée (15 entrées maximum) inspirée des thématiques du projet CATAI. Elle combine **sources académiques** et
références web utiles pour approfondir les domaines abordés :

RÉFÉRNCES BIBLIOGRAPHIQUES

1. Batty, M. (2013). *The New Science of Cities*. MIT Press.
2. Croitoru, A., & Nayak, S. (2014). *Geospatial Data and Smart Cities*. In: Smart Cities and Smart Spaces. Springer.
3. Mitchell, W.J. (1999). *e-Topia: Urban Life, Jim—but Not As We Know It*. MIT Press.
4. Townsend, A. (2013). *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. W. W. Norton & Company.
5. Nunes, M., & Camargo, C. (2017). *Urban Computing and Smart Cities*. In Handbook of Smart Cities. Springer.
6. Allwinkle, S., & Cruickshank, P. (2011). *Creating Smart-er Cities: An Overview*.
7. *Journal of Urban Technology*, 18(2), 1–16. Ratti, C., & Claudel, M. (2016). *The City of Tomorrow: Sensors, Networks, Hackers, and the Future of Urban Life*. Yale University Press.
8. Mora, L., Deakin, M., & Reid, A. (2019). *Smart City Development Paths: Insights from 25 Global Cases*. *Cities*, 92, 222–232.

RÉFÉRENCES WEB ET PROJETS LIÉS

9. [MIT City Science Lab](https://cities.media.mit.edu/)
10. [Smart City Institute – Université de Liège](http://labos.ulg.ac.be/smart-city/homepage/)
11. [Smart Cities Center–Columbia University](https://datascience.columbia.edu/smart-cities)
12. [CASA – University College London](https://www.ucl.ac.uk/bartlett/casa)
13. [CityLab@Inria](https://citylab.inria.fr/)
14. [Programme européen ESPON](https://www.espon.eu/) – projets de données spatiales européennes
15. [Open Geospatial Consortium (OGC)](https://www.ogc.org/) – Normes pour la géovisualisation et l'interopérabilité
16. [MIT City Science Lab](https://cities.media.mit.edu/)
17. [Smart City Institute – Université de Liège](http://labos.ulg.ac.be/smart- city/homepage/)
18. 3. [Smart Cities Center–Columbia University](https://datascience.columbia.edu/smart-cities)
19. [CASA – University College London](https://www.ucl.ac.uk/bartlett/casa)
20. [CityLab@Inria](https://citylab.inria.fr/)
21. [Programme européen ESPON](https://www.espon.eu/) – projets de données spatiales européennes
22. [Open Geospatial Consortium (OGC)](https://www.ogc.org/) – Normes pour la géovisualisation et l'interopérabilité

RAINFALL MODELING AND PREDICTION USING EDGE COMPUTING FOR THE COLOMBIAN ENVIRONMENT

IRENE ARROYO DELGADO1[0000−0003−3014−2395], OSCAR CARRILLO2[0000−0001−5081−1774], AND FRÉDÉRIC LE MOU¨EL3[0000−0002−7323−4057]

1 INDUSTRIAL ENGINEERING, UNIVERSIDAD POLITÉCNICA DE MADRID, SPAIN

IRENE.AD97@GMAIL.COM

2 UNIV LYON, CPE LYON, INSA LYON, CITI, F69621 VILLEURBANNE, FRANCE

OSCAR.CARRILLO@CPE.FR

3 UNIV LYON, INSA LYON, CITI, F69621 VILLEURBANNE, FRANCE

FREDERIC.LEMOUEL@INSALYON.FR

ABSTRACT.

Nowadays the number of devices connected to internet which offer the possibility to collect data is increasing. The interconnectivity of these new sensors favors the creation of sustainable cities, in which the optimization of resources is based on the collected data. These sensors are also a big source of information for forecasting future values.

In this work we present an Edge Computing approach for the analysis and forecasting of rainfall data that is later validated on the CITI Laboratory Youpi Platform. To this end, we built a container image with the necessary tools and libraries to use the time series prediction models SARIMA and Prophet on ARMv7 architectures. A Raspberry Pi 3 node was chosen to evaluate performance on an Edge Computing device.

Colombia was chosen due to its tropical location and its variant geography which present a wide range of historical rainfall data. The data we used to train our models consisted on the historical mesures from sensors deployed in Colombia by the “Instituto de Hidrología, Meteorología y estudios Ambientales de Colombia IDEAM”. In first place, we selected Bucaramanga to study its data sensors and to define the wellsuited parameters for SARIMA and Prophet trend models.

The comparison between them presented a high degree of similarity, offering a good prediction of dry and wet seasons. Thereafter, the SARIMA and Prophet model of Bucaramanga were used to observe its adaptability to the cities of Bogota´ and Medellín, getting a successful outcome at seasonal predictions.

After this estimation of the SARIMA parameters and its analysis offline, a container image was created to simplify and speed up the models implementation in the devices for predicting the two years monthly rainfall for Bucaramanga, Bogota´ and Medellín. The container is available for armv7 architectures, that is usually used for IoT nodes on the Youpi platform.

The proposed model allows to create a network of sensors, with distributed analysis capacity, that improve the prevention of flood or drought emergencies in Smart Cities on Colombia, helping to manage resources for agriculture or prevent catastrophes.

Keywords: Data analytic · Docker · Internet of Things (IoT) · SARIMA · time series prediction · Prophet · Raspberry Pi · Edge Computing.

1. INTRODUCCIÓN

Hoy en día los conceptos de IoT [2] y Smart City [4] están adquiriendo un papel fundamental en la revolución de la Industria 4.0. Según la empresa internacional de investigación y consultoría de tecnologías de la información Gartner, se espera que para 2020 haya más de 20 mil millones de dispositivos de IoT conectados [6].

Un alto porcentaje de dispositivos IoT se encuentran integrados en las llamadas Smart City, su fin es estudiar el comportamiento de las ciudades para establecer medidas que permitan hacer un uso eficiente de los recursos como el transporte e infraestructuras.

La utilización de machine learning y análisis de datos [13] ha sido introducida en las ciudades inteligentes, permitiendo la mejora la calidad de vida e impulsando la economía, al favorecer el uso de predicciones para los modelos de producción y consumo sostenible.

En ellas se deben afrontar nuevos retos de innovación con el fin aumentar la seguridad y privacidad, además de la calidad de los datos que se analizan[8].

1.1 PROPOSICIÓN

A partir de la previsión de aumento de la población se ha acentuado la necesidad de economizar los recursos de los que se dispone, preservando el medio ambiente para las futuras generaciones. Para llevar a cabo una adecuada gestión de los recursos, una de las herramientas claves es la previsión, que se basa en la toma de medidas con el fin de predecir y prepararse para mitigar las catástrofes medioambientales. Una referencia para la prevención de riesgos es la ciudad de Praga, con su estudio de simulación de inundaciones con el fin de establecer protocolos de emergencia para las situaciones más críticas[12].

En este contexto nace el presente trabajo donde se realiza la predicción de la precipitación de la región de Bucaramanga en Colombia, mediante modelos de predicción de series temporales SARIMA y Prophet en nodos de IoT.

Para ambos modelos de predicción, se estudia su portabilidad para la ciudad de Bogotá y Medellín.
La finalidad, es crear una red de sensores puedan actuar de manera autónoma en las estaciones de meteorología y prevengan de comportamientos anómalos basados en la predicción que se realice a partir de los datos históricos. El contenido de la aplicación que se ejecuta en las Raspberry Pi 3 para la realización de las predicciones es:

1. Análisis de los datos.
2. Predicción.
3. Validación de la predicción.

2 METODOLOGÍA

Los métodos utilizados para la estimación de los modelos son SARIMA y Prophet que a continuación se detalla su ajuste.

2.1 SARIMA

SARIMA “Seasonal AutoRegressive Integrated Moving Averages” es un conocido modelo estadístico para analizar series de datos y predecir futuros valores únicamente con la dependencia que existe entre los datos históricos [5,11].
El modelo SARIMA(p, d, q)(P, D, Q, m) se constituye de 7 parámetros, los primeros tres (p,d,q) representan la parte regular del modelo, mientras que (P,D,Q,m) representan la parte estacional.

P, p: término de la componente auto regresiva (AR).
D, d: término de la componente diferencial(I).
Q, q: término de la componente de media móvil (MA).
m: periodicidad de la serie.

Para determinar los parámetros, se debe seguir la metodología BoxJenkins:

1. Identificación y estimación del modelo: Comprobación de la estacionaridad, identificación si procede de estacionalidad y determinación de p y q a partir de las gráficas de autocorrelación y autocorrelación parcial.
2. Determinación de parámetros con algoritmos de cálculo: A través del valor AIC (“Akaike Information Criterion”) junto con el error cuadrático medio (ECM), se obtienen unos parámetros de mayor precisión.
3. Verificación: Se analizan los coeficientes, la bondad de ajuste y se comprueba que los residuos sigan un proceso de ruido blanco.
4. Predicción: Una vez que todas las condiciones anteriores se cumplen, el modelo esta listo para realizar la predicción, bajo la hipótesis que los valores futuros están relacionados con el pasado. Se realiza una predicción de 24 meses.

2.2 Prophet
Facebook publicó en 2017 una herramienta llamada Prophet. Es una librería básica de código abierto que permite hacer modelos y predicciones de series temporales. Se basa en un modelo de regresión aditivo o también llamado “curve fitting” que se representa por la siguiente ecuación.

$$y(t) = g(t) + s(t) + h(t) + Et \quad (1)$$

g(t): función de tendencia de crecimiento logarítmico o linear para modelos con cambios no periódicos en la serie.

s(t): función de cambios periódicos y estacionales.
h(t): efecto de las vacaciones y eventos.
Et: término de errores no modelizados.

En el modelo Prophet, se configura la estacionalidad anual, los valores de saturación, el porcentaje de valores donde aplicar los puntos de cambio de tendencia y el intervalo de confianza. Mientras que los parámetros internos del modelo se autoajustan para cada serie. Con el objetivo de la validación del modelo, Prophet incorpora para el diagnóstico la validación cruzada (“cross validation”) y las métricas de rendimiento (“metrics performance”) [14,9].

3. IMPLEMENTACIÓN

Los modelos se han determinado offline a partir de los datos de Bucaramanga y posteriormente, se han exportado a un nodo de IoT para la predicción de datos futuros.

3.1 ANÁLISIS DE LOS DATOS

Para la realización de los modelos, se necesitaba en primer lugar, una base de datos fiable y segura. Para dicha aplicación, se analizó la precipitación acumulada mensual. La fuente de datos fue el Instituto de Hidrología, Meteorología y estudios Ambientales IDEAM [7]. Los sensores proporcionaban la precipitación acumulada diaria en mm de las estaciones de la región de Bucaramanga, a los que se les debió pasar un filtro para descartar los datos fuera de los límites naturales. El filtro máximo fue seleccionado en base a cuál había sido la precipitación máxima diaria de los últimos años en la zona donde se situá la estación, mientras que el filtro mínimo fue cero mm por obviedad.

3.2 MODELO SARIMA PARA BUCARAMANGA

Para la elección de los siete parámetros del modelo SARIMA, fue necesario seguir los pasos anteriormente mencionados con el fin de obtener el modelo para Bucaramanga. En primer lugar, se comprobó la estacionaridad, para ello se usó el método de DickeyFuller. Al aplicar la función a la serie anteriormente tratada, se obtuvo que el método pasaba para todos los intervalos de confianza estudiados, pero la media y varianza no seguín una tendencia estacionaria. Por lo que se decidió aplicar diferencia de orden uno.

La diferencia de orden uno para el parámetro regular aportó una media y varianza estacionarias, por lo tanto, se adoptó para el modelo d=1 y D=0. Posteriormente, con el fin de determinar manualmente una aproximación de los parámetros del modelo, se analizó la ACF y PACF de la serie. Estas funciones presentaban una forma de abanico que se completaba en un periodo aproximadamente 12, m=12. Por lo que, se observó que poseía componentes estacionales y un ciclo anual. Al ser la modelización de series meteorológicas tan compleja, los parámetros restantes se obtuvieron

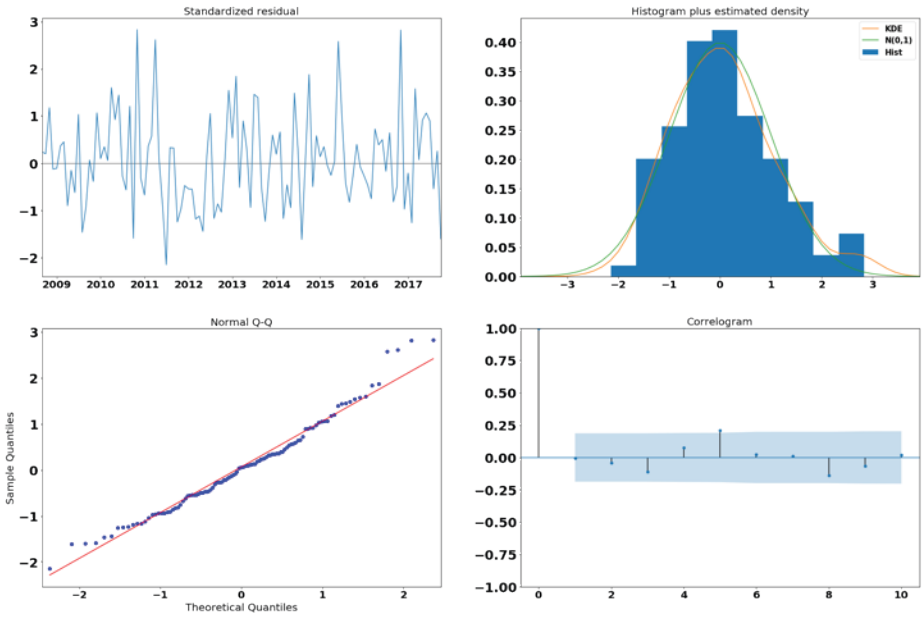
con el método AIC. Se procedió a encontrar el modelo con menor AIC y menor ECM. Con esta finalidad, se realizó un “grid search” donde se obtuvieron los tres modelos con parámetros entre [0,3] con menor AIC. De los cuales, se seleccionó como mejor modelo el que menor error disponga.

Para la validación del modelo, se determinanó los parámetros de ajuste entrenando el modelo con el 90% de los datos y posteriormente verificando con el 10% restante. A partir de la grid search se obtuvo que el modelo elegido es SARIMA (2, 1, 3) (2, 0, 3, 12) al ser el que menor error cuadrático medio presentaba. Los modelos SARIMA ofrecen la posibilidad de estudiar el peso e importancia de sus parámetros.

Para este modelo, se estudió que ar.S.L24, el parámetro correspondiente a P de orden 2, tenía un valor P>z muy alto y un coeficiente pequeño, es decir tenía poco peso y una importancia despreciable en la predicción por lo que se podía eliminar simplificando el modelo. El modelo definitivo fue SARIMA (2, 1, 3) (1, 0, 3, 12). El cual fue validado al comprobar que sus residuos seguían una distribución de ruido blanco, es decir los residuos seguían un proceso de distribución aleatorio,

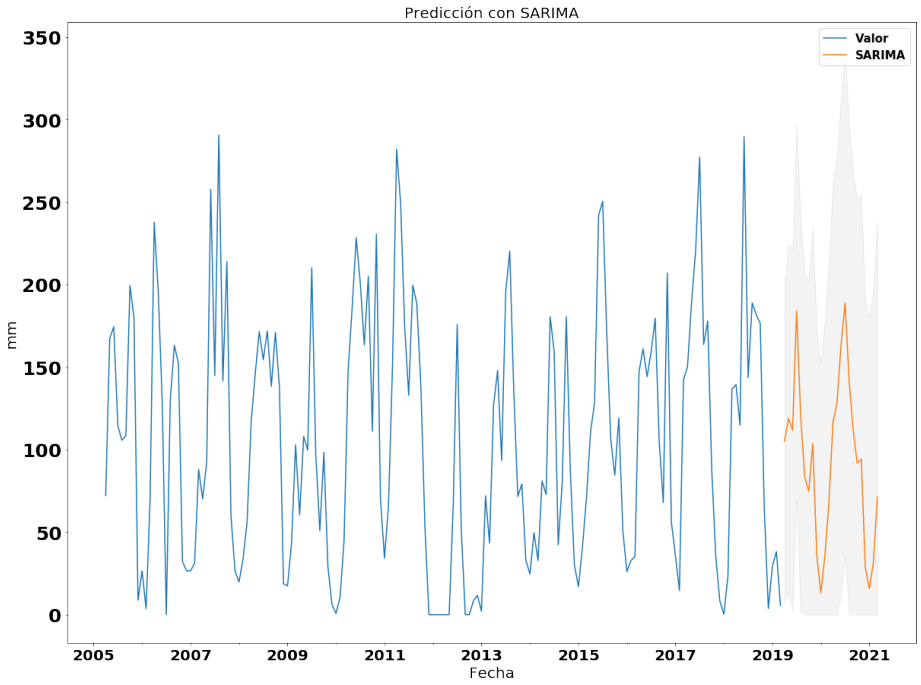
Figura 1, con una correcta aproximación de la predicción para el 10% de los datos usados en el previo entrenamiento y un buen ajuste a los datos reales de la serie.

Fig. 1. Residuos del modelo SARIMA



Con el modelo ya establecido se realiza la predicción para 24 meses, Figura 2, dos años, con un intervalo de confianza del 95%. Esta predicción sigue los estándares marcando las temporadas de lluvias y secas.

Fig. 2. Predicción para 24 meses en Bucaramanga con SARIMA



3.3 MODELO PROPHET PARA BUCARAMANGA

En Prophet para Python, se analizan los puntos de cambio de tendencia en el 80% de los datos, para una mayor exactitud en el presente trabajo se aumentó hasta el 90% dejando espacio para la proyección de la tendencia, pero analizándose un mayor número de puntos. Por otro lado, se aumentó la flexibilidad de la tendencia, añadiendo más prioridad a los puntos de cambio "changepoint" a un 0.7 debido a que en dicho modelo es necesario ajustarse a las tendencias de los años posteriores.

Además, se eliminaron los datos atípicos y en su lugar, el modelo estimaría que valores corresponderían según la tendencia en ese intervalo de tiempo. Esta propiedad se utilizó con el fin de que no se tuviesen en cuenta los datos de 200607, 201112 201206, 201209 y 201210 que estarían definidos a 0 mm, para la estación de Paramo del Almorzadero, Bucaramanga. Otro aspecto a destacar ha sido la fijación de la saturación mínima a 0, debido a que no es físicamente posible obtener lluvia negativa. La predicción para 24 meses fue efectuada con el modelo Prophet siguiendo una tendencia realista, cumpliendo los objetivos con un intervalo de confianza del 95% como se puede corroborar en la Figura 3 .

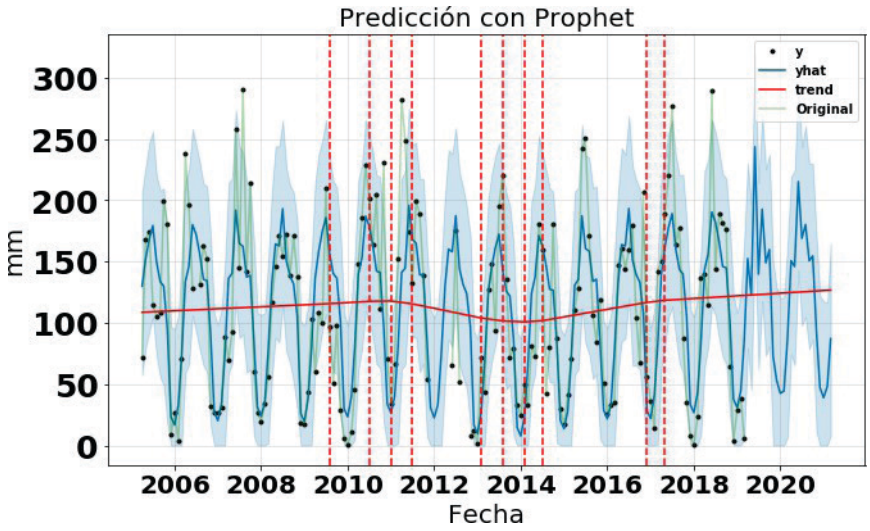


Fig. 3. Predicción para 24 meses en Bucaramanga con Prophet

4. MODELO EN NODOS IOT

A fin de ejecutar los modelos creados en Jupyter notebook es necesario tener instaladas las librerías específicas para cada proceso, además del propio Jupyter notebook y Python en la Raspberry Pi 3. Este no es un dispositivo IoT, sin embargo, cuenta con una arquitectura propia de una amplia cantidad de ellos, por lo que se utilizan para realizar simulaciones de comportamiento. El proceso de descarga de cada paquete es lento al tener la Raspberry Pi una conexión Ethernet más lenta que un ordenador personal al compartir el bus USB de la placa para realizar la conexión. De manera que se agilice el tiempo de descarga y su portabilidad, se creó una imagen Docker con todo lo necesario, con el fin automatizar su implementación.

Se implementó la imagen Docker para arm/v7 desde el Escritorio Docker a partir de compilación cruzada. Esta funcionalidad fue anunciada el 24 de abril de 2019 por lo que es una herramienta todavía en versión beta. Docker Desktop o Escritorio Docker, ha incorporado esta funcionalidad debido a la gran expansión de los dispositivos de IoT que presentan estas arquitecturas ARM [10].
5. Resultados y Conclusiones
En primer lugar, se realizó un estudio de los diferentes sensores del Instituto de Hidrología, Meteorología

y estudios Ambientales IDEAM para la ciudad de Bucaramanga. Se estudiaron catorce sensores de los cuales, al aplicar el filtro de corrección de datos y representarlo gráficamente, se observó que en todos los sensores menos en uno había grandes intervalos de datos atípicos o intervalos de datos escuetos, por lo que se llegó a la conclusión de que la única estación correcta para la realización de predicción de series temporales era Paramo del Almorzadero, que se encuentra situada a 46 km de Bucaramanga.

Por otra parte, al realizar un estudio similar posteriormente para Bogotá y Medellín se pudo observar nuevamente estos intervalos prolongados de datos atípicos y siniestros en las estaciones. De esta manera se afirmó que las estaciones de IDEAM en las ciudades de Bucaramanga, Bogota´ y Medellín presentaban un elevado número de sensores con datos incorrectos.

Posteriormente, se estudió la serie temporal de Páramo del Almorzadero para determinar un modelo SARIMA y Prophet que se ajustase correctamente y proporcionase una predicción fiable para la ciudad de Bucaramanga. Para el modelo SARIMA, se utilizó el método AIC combinado con el error cuadrático medio, obteniendo como resultado que el mejor modelo era SARIMA (2, 1, 3) (2, 0, 3, 12). Tras un análisis de sus coeficientes se observó que se podía reducir el orden de AR(P) sin alterar el resto de los parámetros, por lo que se obtuvo como modelo final SARIMA(2, 1, 3)(1, 0, 3, 12).

Para el modelo con Prophet, los parámetros son autoajustables, por lo que solo se identificaron las zonas con datos atípicos.

Al comparar las predicciones de la Figura 4 , se observó que presentaban los mismos cambios de estaciones, época lluviosa entre abrilmayo y septiembreoctubre, y temporada seca en los intervalos restantes, siendo más relevante enero donde la precipitación es mínima. Por lo tanto, comparadas con el clima en esta región y sus estaciones, la predicción presenta un estudio realista de los periodos estivales y con respecto a la predicción de valores, ambas presentan un elevado

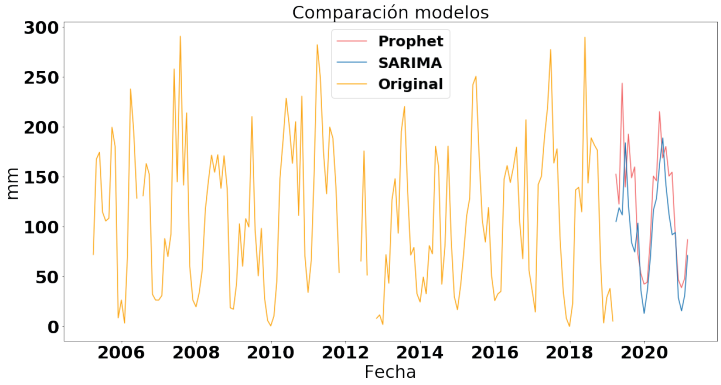


Fig. 4. Comparación modelos SARIMA y Prophet para Bucaramanga

valor aproximado entre ellas. Se concluye, que ambos modelos representan una predicción fiable, bien si a número de mm de precipitación acumulada mensual no es totalmente exacto, al incluir un nivel de confianza del 95%, se utilizará para estudiar las distintas temporadas y próximas tendencias con gran exactitud.

A continuación, se buscaba comparar como ajustan los modelos anteriormente seleccionados a otras series temporales, para comprobar si los modelos que se han seleccionado eran generalizables para otras ciudades con meteorología parecida. Se estudió para las ciudades de Bogotá y Medellín.

En el ajuste para Bogotá, se estudió el sensor de Páramo de Chingaza, situado a 38 km de Bogotá. La predicción de las estaciones de la Figura 5 coincide

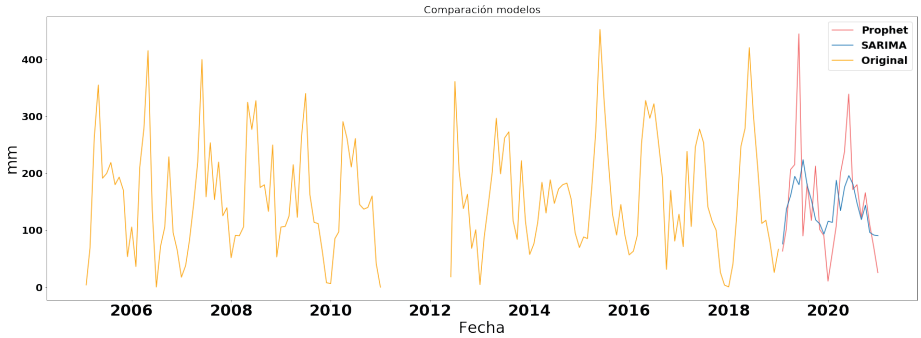


Fig. 5. Comparación modelos SARIMA y Prophet para Bogota´

con las de esta región, bien si a valor numérico para el caso de SARIMA es inferior al de Prophet en la temporada húmeda y mayor en la temporada seca. Se concluye que los modelos se ajustan satisfactoriamente a nivel de estaciones y respecto a una mayor exactitud de la precipitación mensual hay que incluir el intervalo de confianza. En el ajuste de Medellín, se estudió el sensor de Aragón a 60 km de Medellín La predicción que se realizó en la Figura 6 sigue los estándares

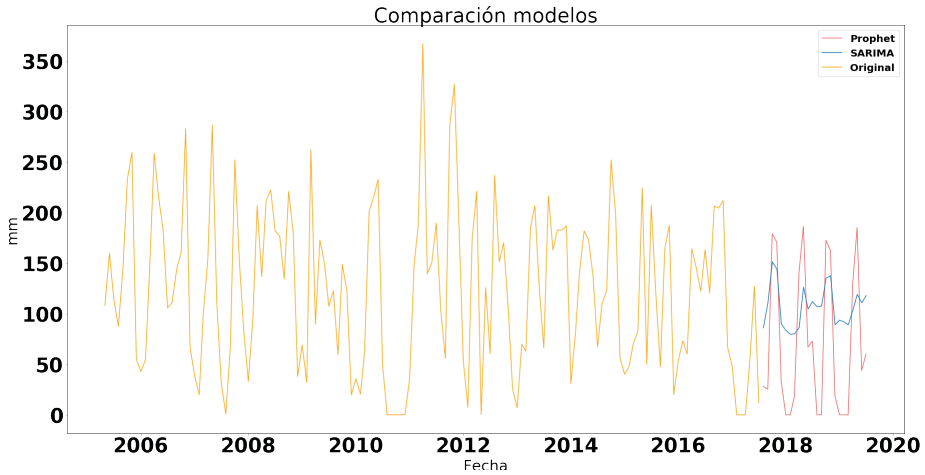


Fig. 6. Comparación modelos SARIMA y Prophet para Medellín

respecto a las estaciones húmedas y secas. Sin embargo, respecto a los valores de precipitación mensual el modelo SARIMA presenta unos niveles mínimos elevados. Se concluye que ambos modelos representan correctamente las temporadas de precipitación, pero a nivel de exactitud Prophet es más adecuado. Con el fin de exportar los modelos en Jupyter Notebook a los dispositivos IoT.

Se creó una imagen Docker dynamid/rainforecastiarroyo con la cual se facilita y agiliza el proceso. Al comparar los tiempos de ejecución, el proceso en la Raspberry Pi 3 fue más lento que en el ordenador personal de un orden de 7.4 veces superior para SARIMA y 8.2 veces superior para Prophet. A pesar de su ejecución más costosa en tiempo, la Raspberry Pi 3 utilizada representa un dispositivo que soporta las herramientas necesarias para dichos modelos con un rendimiento aceptable.

Consecuentemente, se ha demostrado la factibilidad de ejecutar los modelos de predicción de lluvia en arquitecturas armv7 que hasta el momento no se había realizado debido a que los nodos no disponían de la capacidad de compilación de las herramientas de análisis de series de datos, que gracias a la compilación cruzada ha sido posible que se compilen en una arquitectura más eficiente.

Como síntesis final, se puede decir que se ha conseguido obtener los modelos de SARIMA y Prophet, para realizar una aplicación en Jupyter Notebook capaz de realizar predicciones de la precipitación exportables a una red de nodos IoT con arquitectura armv7 gracias a la imagen Docker suponiendo un incremento en el tiempo respecto del ordenador personal debido a la arquitectura y componentes de la Raspberry. Los modelos elegidos para Bucaramanga son adecuados a su vez para Bogotá y Medellín para indicar las estaciones húmedas y secas, además de la tendencia para los próximos dos años [1].

Este trabajo supone un aporte en la transformación de las ciudades a Smart Cities con la capacidad de predecir y poder simular posibles escenarios de precipitación meteorológica y mitigar los efectos adversos de un cambio de tendencia gracias a herramientas de machine learning y data analytic. Con este proyecto, se abren las puertas a nuevas investigaciones para crear una plataforma de pruebas con una red de recogida de datos meteorológicos en tiempo real como la existente en Birmingham [3].

REFERENCES

1. Arroyo Delgado, I.: Predicción de precipitaciones con dispositivos IoT mediante ana´lisis de series temporales. B.S. Thesis, Universidad Politécnica de Madrid (2019)
2. Atzori, L., Iera, A., Morabito, G.: The Internet of Things: A survey. Computer Networks 54(15), 2787–2805 (2010). https://doi.org/10.1016/j.comnet.2010.05.010, http://linkinghub.elsevier.com/retrieve/pii/S1389128610001568
3. Chapman, L., Muller, C.L., Young, D.T., Warren, E.L., Grimmond, C.S., Cai, X.M., Ferranti, E.J.: The Birmingham urban climate laboratory: An open meteorological test bed and challenges of the Smart city. Bulletin of the American Meteorological Society 96(9), 1545–1560 (2015). https://doi.org/10.1175/BAMSD1300193.1
4. Ching, T.Y., Ferreira, J.: Smart Cities: Concepts, Perceptions and Lessons for Planners, pp. 145–168. Springer International Publishing, Cham (2015). https://doi.org/10.1007/9783319183688 8
5. de la Fuente Ferna´ndez, S.: Series Temporales: Modelo Arima. Universidad Autónoma de Madrid p. 53 (2016), http://www.estadistica.net/ECONOMETRIA/ SERIESTEMPORALES/modeloarima.pdf
6. Hung, M.: Leading the IoT. Gartner insights on how to lead in a connected world. Gartner (2017), http://gartner.com/imagesrv/books/iot/iotEbook% 5Fdigital.pdf
7. IDEAM: Instituto de Hidrología, Meteorología y estudios Ambientales, http:// www.ideam.gov.co
8. Naphade, M., Banavar, G., Harrison, C., Paraszczak, J., Morris, R.: Smarter cities and their innovation challenges. Computer 44(6), 32–39 (2011). https://doi.org/10.1109/MC.2011.187
9. Nishida, K.: An Introduction to Time Series Forecasting with Prophet in Exploratory (2017), https://blog.exploratory.io/anintroductiontotimeseriesforecastingwithprophetpackageinexploratory129ed0c12112
10. Parko, A.: Building MultiArch Images for Arm and x86 with Docker Desktop (2019), https://engineering.docker.com/2019/04/multiarchimages/
11. Peña, D.: Ana´lisis de Series Temporales. Alianza (2005)
12. Rothkrantz, L.J.: Flood control of the smart city Prague. 2016 Smart Cities Symposium Prague, SCSP 2016 pp. 1–7 (2016). https://doi.org/10.1109/SCSP.2016.7501043
13. de Souza, J.T., de Francisco, A.C., Piekarski, C.M., do Prado, G.F.: Data mining and machine learning to promote smart cities: A systematic review from 2000 to 2018. Sustainability (Switzerland) 11(4) (2019). https://doi.org/10.3390/su11041077
14. Taylor, S.J., Letham, B., Taylor, S.J., Letham, B.: Forecasting at Scale Forecasting at Scale. The American Statistician 72(1), 37–45 (2018). https://doi.org/10.1080/00031305.2017.1380080

ANALYSIS OF URBAN INFORMATION OF COLOMBIA ON WIKIDATA

JOHN SAMUEL1,2[[0000–0001–8721–7007] AND OSCAR CARRILLO1,3[0000–0001–5081–1774]

1 CPE LYON, UNIVERSIT´E DE LYON, LIRIS, UMR 5205, VILLEURBANNE, FRANCE

2 LIRIS, UMR 5205, VILLEURBANNE, FRANCE

3 CITI LYON, INSA DE LYON, LYON, FRANCE

{JOHN.SAMUEL,OSCAR.CARRILLO}@CPE.FR

ABSTRACT.

Wikidata started in 2012 is a free, open, multilingual, col-laborative, linked and structured knowledge base. During the last couple of years, Wikipedia communities of different languages have started exploring the use of Wikidata as a central store of information, where the community members can directly add structured data on Wikidata to be later be used by all the language editions of Wikipedia. However, making it a central store has several challenges, especially when key in-formation like administrative heads, population etc. are outdated.

In this article, we will first consider the available urban data related to some of the main cities of Colombia on Wikidata and see what type of infor-mation are currently contributed by the community members. We will then compare the historic evolution of these contributions with respect to the size of the cities in terms of the area, population, tourists as well as the influence of important moments and events on the recent history of Colombia.

We will also focus on the frequency and recency of the con-tributions. The analysis of these data is important since it will ensure that relevant government agencies, Wikidata community members and tourism organizations work together to keep the information up to date. We will also discuss how our study can be further explored by other cities as well as countries for ensuring timely information on Wikidata and associated Wiki-media projects.

Keywords: Wikidata · Urban Data · Collaboration.

INTRODUCTION

The world is increasingly referred to as global village [3], thanks to the growing inter-connectivity among different cities, especially with the advances of trans-port infrastructure as well as the internet usage.

It is much easier to know in-formation on remote towns and villages now, than it was a couple of decades ago, thanks to the growing availability of linked open data [1] which encourages autonomous institutions to not only publish their data with open licences but also link with each other. Websites like Wikipedia provide a platform where users across the world can write articles on various topics, including countries, capitals, towns, villages etc. Wikipedia also allows these articles to be written in almost more than 300 multiple

languages. These multilingual articles are linked to each other, thereby letting users easily switch and see other languages versions of the information available for a given subject. However, much of the content on Wikipedia is unstructured and several efforts have been made to extract relevant information from Wikipedia [9] including Wikipedia infoboxes [1, 2].

With the advent of Wikidata in 2012 [8], multilingual contributors can make contributions in the form of triples (subject item-property-value statements). Thus to state that Bogot´a is the capital city of Colombia, the contributor can state Q739-P36-Q2841, where Q739 and Q2481 are the identifiers of Colom-bia (subject) and Bogot´a (value) on Wikidata and P36 is the identifier of the property 'capital'. Wikidata (https://www.wikidata.org) with its single domain website address, unlike Wikipedia (a multi-subdomain website) has the advan-tage that users can change the language settings and can see the facts related to a given subject in any supported language (from a list of more than 300 languages) and also contribute in their local languages. Maintaining up-to-date information on all the multilingual Wikipedia sites is challenging and using a central multilingual site like Wikidata for recent information may be a possible way forward. Many language Wikipedias like Catalan, Basque and even English are now exploring to use data from Wikidata to enrich the Wikipedia infoboxes.

Wikidata is a collaborative website like Wikipedia and hence it's extremely important to understand how and what type of facts are entered by the users. Some of these facts evolve during time. Take for example, values like popula-tion of countries, cities change over time. It's therefore important to monitor whether articles (called items on Wikidata) on different subjects contains the latest information. Another major problem with collaborative sites is vandalism [4] and Wikidata also suffers from it, especially in the form of label or description changes[7, 6].

Wikidata is also increasingly being used as a knowledge hub [5] for other knowledge organisation systems. With the initial goal of supporting other Wiki-media projects including Wikipedia, it is now linked to other databases, with the use of external identifiers, thereby allowing its users to verify and also have additional information from these systems. However, maintaining a central site for facts means ensuring the latest up-to date information of all the facts.

In this article, our goal is to look at the current information of different cities of Colombia on Wikidata and see what type of information are currently contributed by the community members like the different properties used by urban cities, the use of external identifiers, use of images, edit reverts, article lengths (number of statements) etc.

Section 2 briefly presents the cities that we took into consideration. Analysis of the information of these cities are done in section 3. We discuss our results in section 4 and briefly present how our work can be further extended to other cities. Finally, we conclude our article in 5.

2. CITIES OF COLOMBIA

We take into consideration the different departmental capitals, intermediary cap-itals, touristic cities, some villages and towns of interest. Table 2 gives a detailed information about our focus categories and the associated list of cities. We ob-tained the 20 main cities of Colombia using the Spanish language Wikipedia template Principales ciudades de Colombia4. We also took into account some of the towns in conflict. In total, we are considering 40 towns, villages and cities.

In the table, we have also given the Wikidata identifiers of each of these towns in parentheses. In the

following section, we will explore the statements used to describe them and the associated multilingual articles.

S.No.	Focus category	List of cities
1.	Main cities of Colombia [5]	Bogota´ (Q2841), Medell´ın (Q48278), Cali (Q51103), Barranquilla (Q62823), Cartagena (Q657461), Soledad (Q767071), Cúcuta (Q216847), Soacha (Q1011151), Ibagu´e (Q222755), Bucaramanga (Q243766), Villavicencio (Q749224), Santa Marta (Q209016), Bello (Q816024z), Valledupar (Q376903), Pereira (Q51111), Buenaventura (Q996581), San Juan de Pasto (Q320015), Manizales (Q235190), Monter´ıa (Q852725), Neiva (Q638260)
2.	Important departmental capitals	Bogota´ (Q2841), Barranquilla (Q62823), Cali (Q51103), Bucaramanga (Q243766), Medell´ın (Q48278)
3.	Intermediary capitals	Armenia (Q328518), Manizales (Q235190) and Monter´ıa (Q852725)
4.	Touristic cities	Cartagena (Q657461), San Andr´es (Q134678), Providencia (Q681111), Leticia (Q214913), Puerto Nariño (Q767738), Santa Marta (Q209016) and El Cocuy (Q1655984)
5.	Villages	Villa de Leyva (Q1409503), San Gil (Q1294128), Barichara (Q1576773) and Zipaquira (Q205429)
6.	Small towns	Mocoa (Q579803), Villavicencio (Q749224), Soledad (Q767071), Pasto (Q320015) and Neiva (Q638260)
7.	Towns in Conflict	Cu´cuta (Q216847), Medellın,Arauca (Q626543), Paraguacho´n (Q6060938), Corinto (Q2236398), Piendamo´ (Q2433349), Santander de Quilichao (Q1093175), Caldono (Q1391900), Jambalo´ (Q1525285), Puerto Tejada (Q1256377), Torib´ıo (Q1870972)

4 https://es.wikipedia.org/wiki/Plantilla:Principales ciudades de Colombia

We make use of the Wikidata SPARQL endpoint6 and Wikidata Mediawiki API7 to collect the relevant information for our analysis.

3. URBAN INFORMATION OF COLOMBIA ON WIKIDATA

Table 3 gives a list of all the properties (total 96) currently used by the con-tributors to describe the cities above. It also shows the property identifiers and the associated data types. We see the use of 9 out of 17 supported data types. The table also shows the use of 36 external identifiers to various external data sources including national libraries and encyclopedic entries.

Figure 1 shows the number of statements used by the selected cities. We see the higher number of statements for Bogot´a, Medell´ın, Cali and Cartagena. In Figure 2, we explore the number of distinct properties used by the different cities. We see Bogot´a using 75 (out of 96) distinct properties.

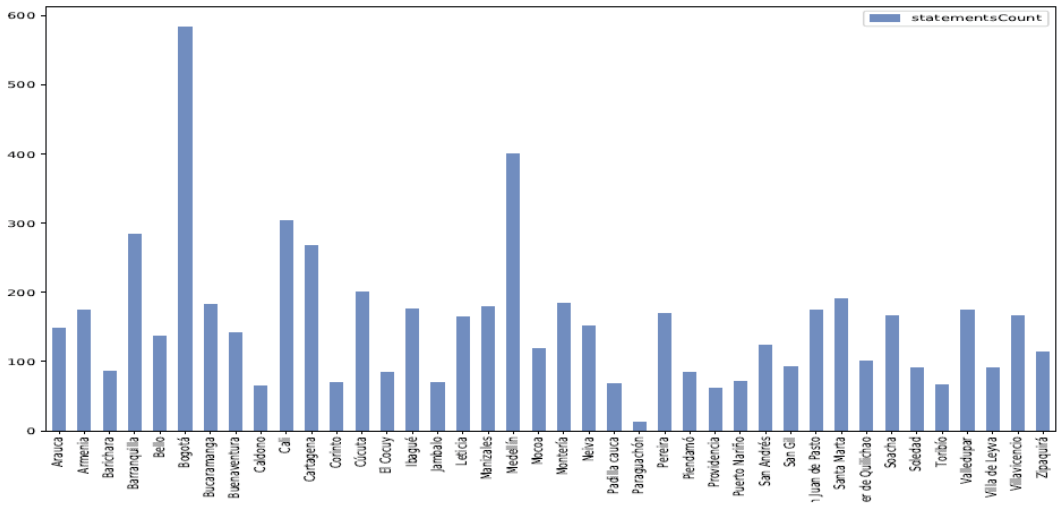


Fig. 1. Number of statements for different cities

Property values can be multi-valued. We compare the number of distinct properties versus the number of statements in Figure 3.

It's very important that all of these statements have associated references. Figure 4 shows the number of references used on the different Wikidata items. Unlike previous figures, Medell´ın is taking the lead. We compare the number of statements and the number of references in Figure 5. We see that in a majority of the cities, the number of references do not even cross 50%.

Next in Figure 6, we look at the number of Wikidata items that make use of the above cities in their statements, i.e., as a property value in the subject-property-value triplet. Here, Bogot´a with 7048 inbound links is significantly ahead of the next city Medell´ın with 2659 links. Similar observation can be found

6 https://query.wikidata.org/

7 https://www.wikidata.org/w/api.php

ANALYSIS OF URBAN INFORMATION OF COLOMBIA ON WIKIDATA

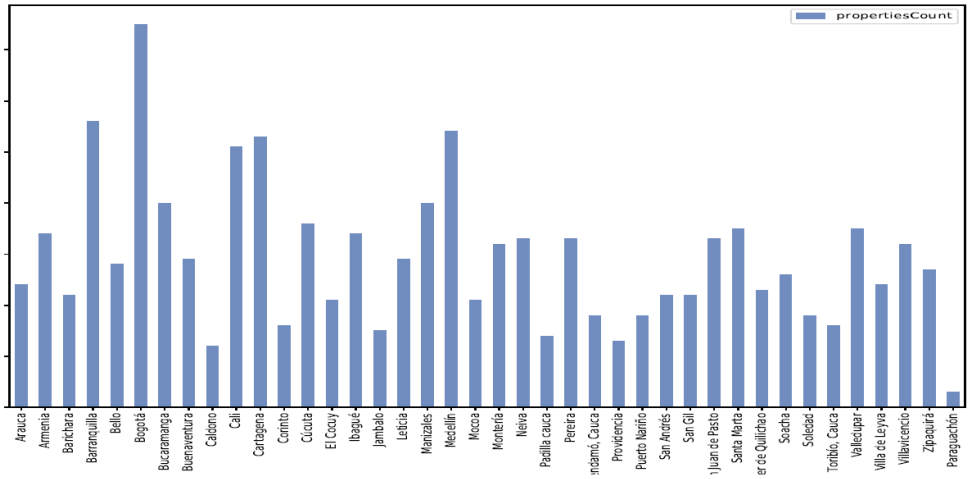


Fig. 2. Number of distinct properties for dierent cities

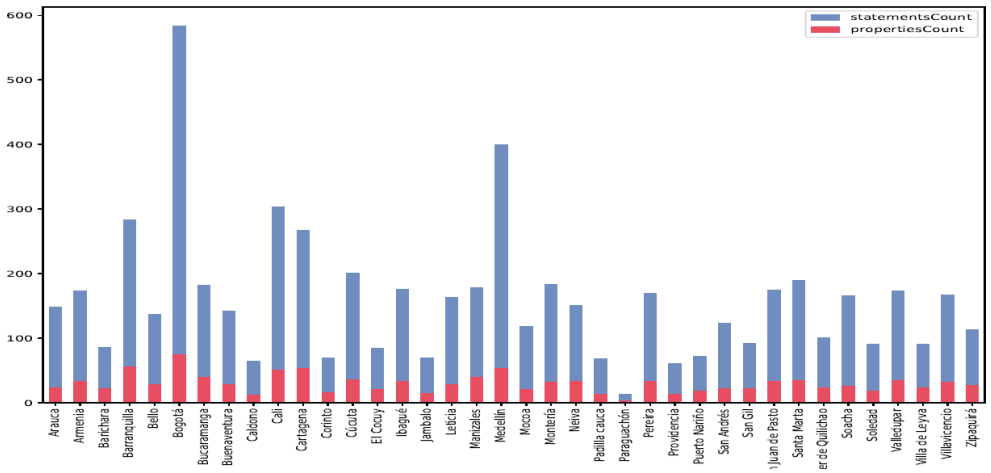


Fig. 3. Number of statements versus number of distinct properties for dierent cities

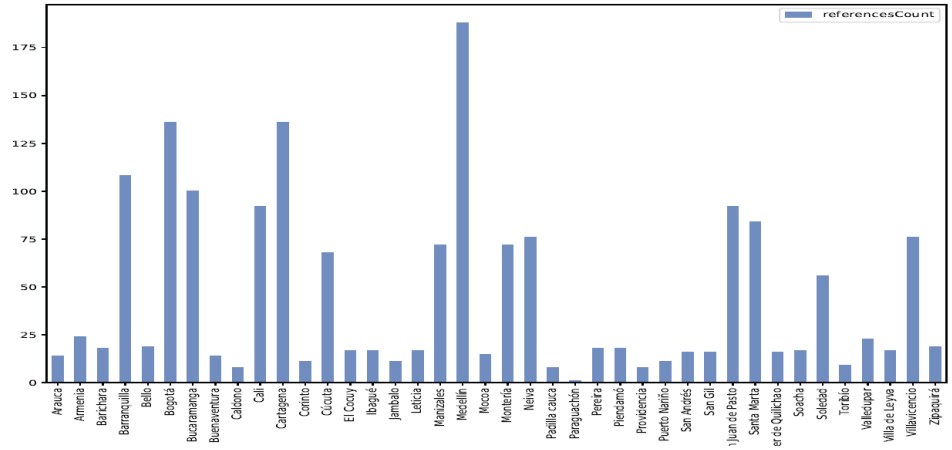


Fig. 4. Number of references for dierent cities

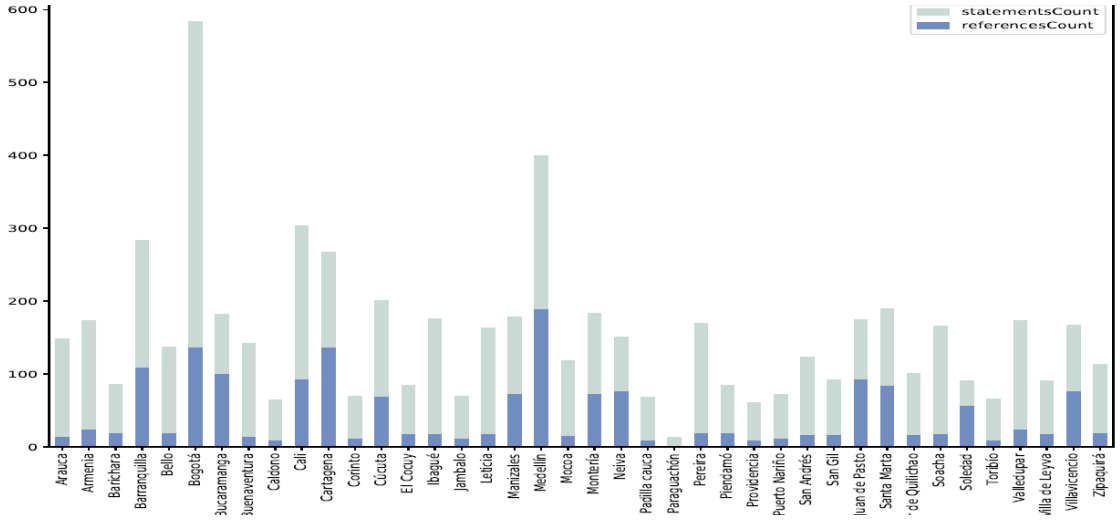


Fig. 5. Number of statements versus number of references for different cities in the number of multilingual

Wikimedia articles (on several projects including Wikipedia, Wikivoyage, Wikispecies etc.) of these cities with Bogot´a taking the lead followed by Medell´ın
(See Figure 7 and Figure 8).

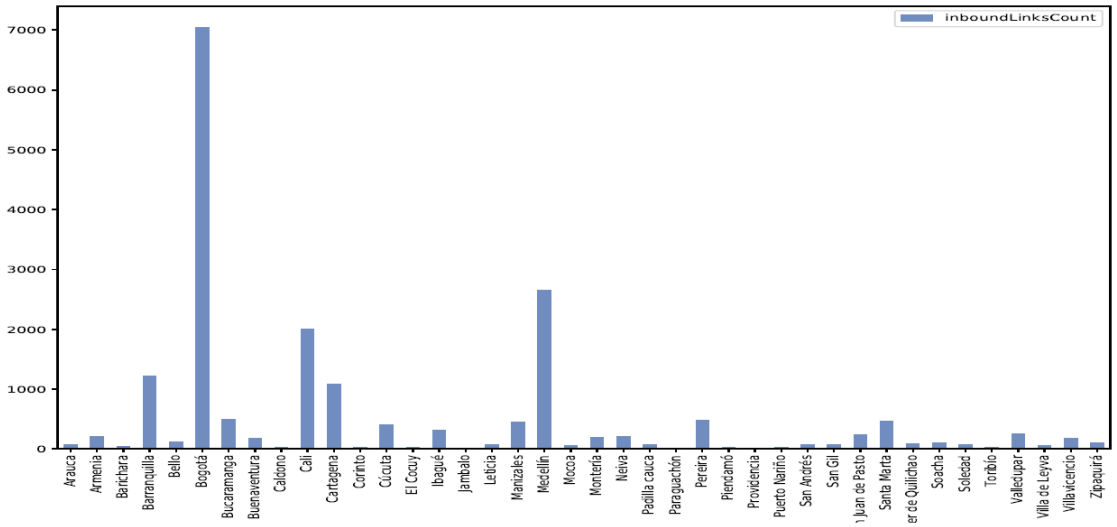


Fig. 6. Number of inbound-links for different cities

We also look at the most commonly used languages by these Wikimedia articles in Figure 9. It may not be surprising to see English taking the lead, followed by Spanish, French, Portuguese etc.
As discussed above, thanks to the use of external identifiers by Wikidata items, users can verify information from other external sources. In Figure 10, we see the number of external identifiers used by the cities.
Finally we take a look at the multilingual labels, descriptions and aliases on Wikidata in Figure 11 first by comparing them and then individually analyzing them in Figure 12.

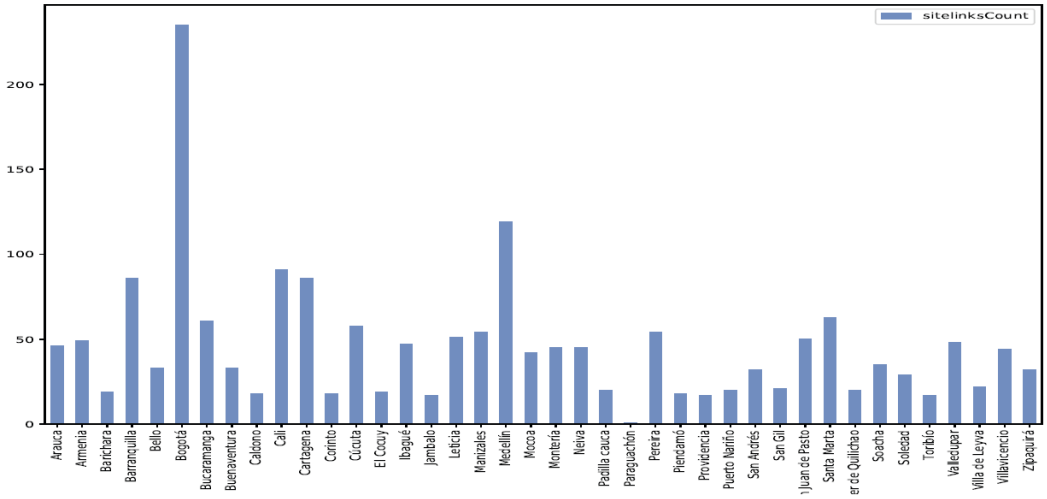


Fig. 7. Number of Wikimedia articles for dierent cities

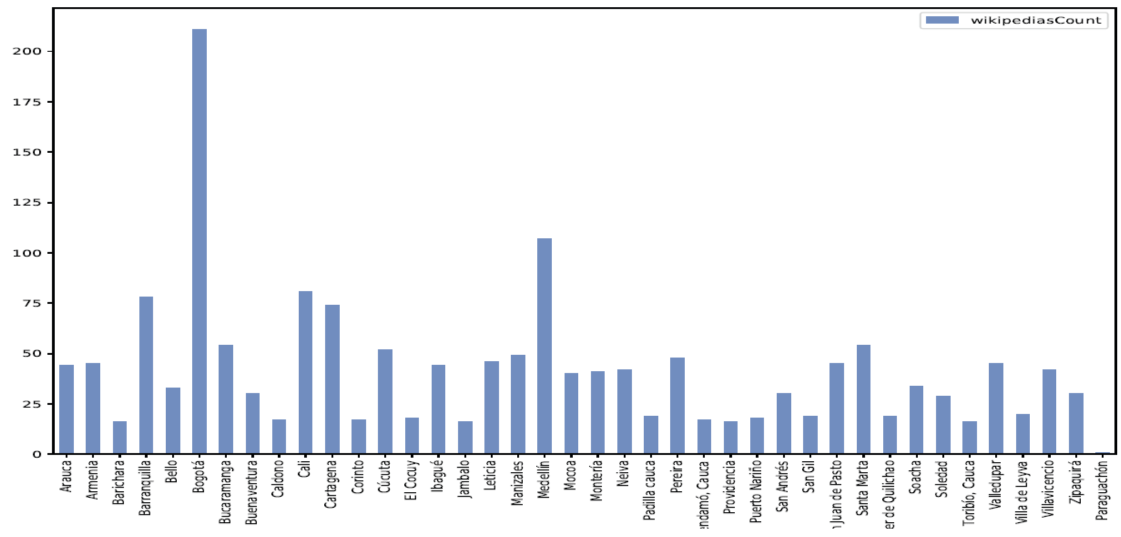


Fig. 8. Number of Wikipedia articles for dierent cities

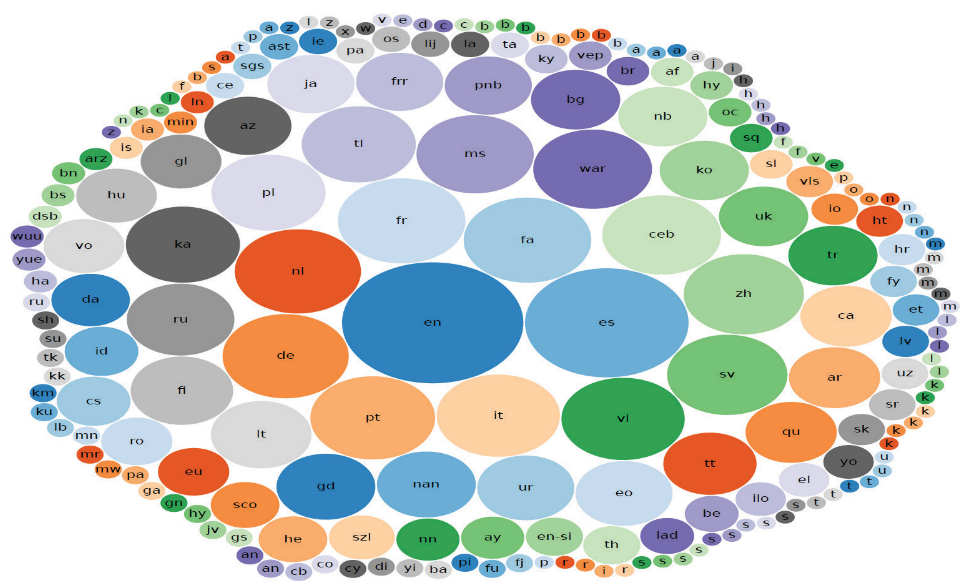


Fig. 9. Bubble graph on the number of articles of the selected cities in different lan-guages

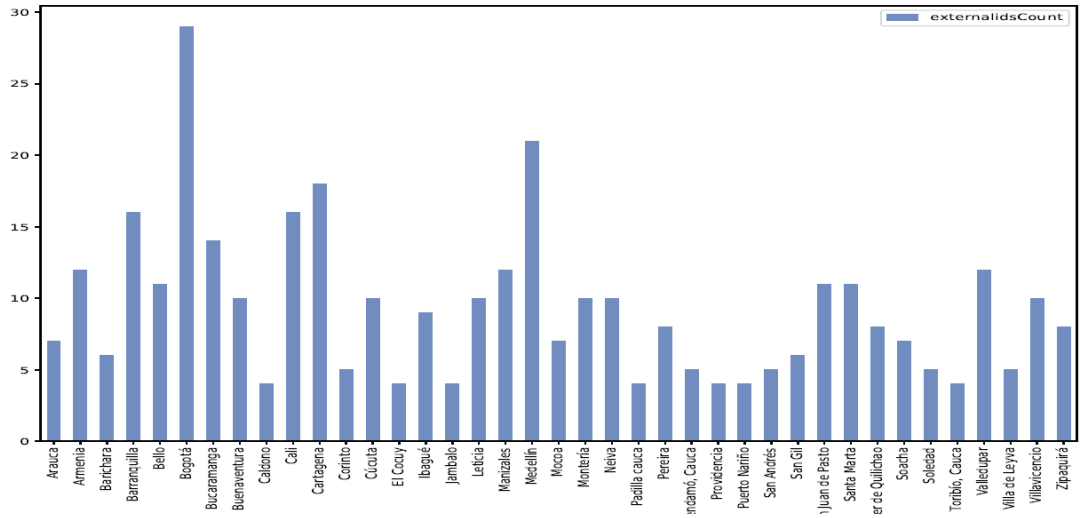


Fig. 10. Number of external identifiers for different cities

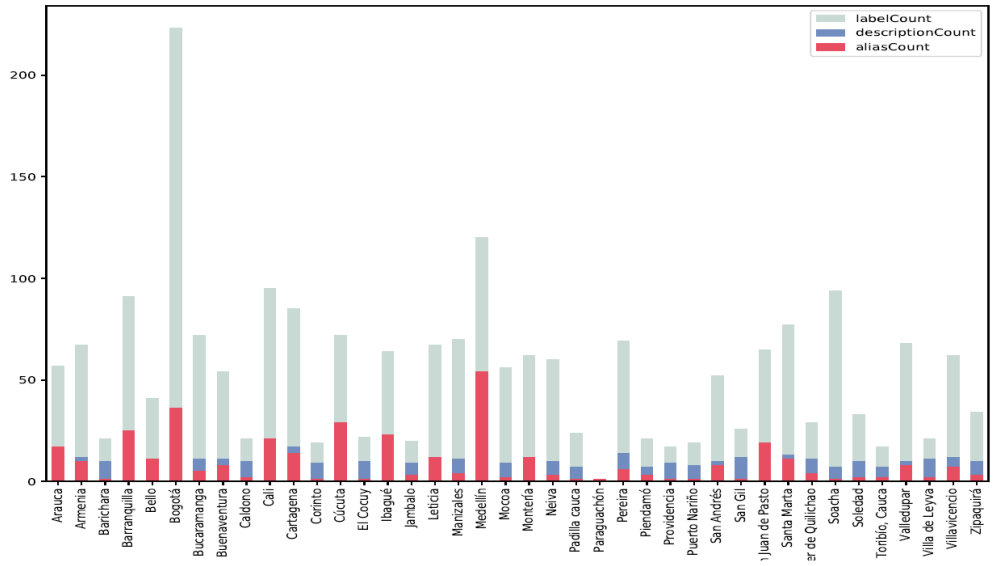


Fig. 11. Number of multilingual labels, descriptions and aliases of different Colombian cities

Fig. 12. Bubble graph on the number of multilingual labels, descriptions and aliases respectively for the selected cities

Current usage of images for these cities is also not complete. Only 33 of the above cities have an associated image (i.e., make use of the property P18). Also approximately 300 of the edits were reverted (detected by making use of ”undo”, ”reverted”, ”Undid”) pointing to possible vandalism attempts.

DISCUSSION

We see the detailed statistics on translation of labels, description and aliases in Table 4. Statistics related to the number of statements, references, inbound Links and Wikimedia articles are presented in Table 4.

It’s interesting to observe that four major cities: Bogot´a, Medell´ın, Carta-gena, Cali have significantly good number of statements and references. 96 dis-tinct properties are used by the selected Colombian cities, towns and villages. Bogot´a uses 75 distinct properties, compared to Paris (131), London (109), New York (108) and Berlin (129). No property using the Wikidata datatype GeoShape is currently being used. A total of 36 external identifiers to various external data sources including national libraries and encyclopedic entries are currently used.

Bogot´a has 29 distinct external identifiers, compared to Paris (58), London (57), New York (50) and Berlin (59).

It’s interesting to compare the statistics between different cities in an inter-active manner so that new contributors can explore how to improve different items. WikiProvenance8 provides such a limited feature. However, it’s unable to compare a large number of cities as explored in this article.

CONCLUSION

The major potential of Wikidata is its multilingual and structured nature that can be used to build personal assistants that can respond to user queries in their local languages.

However, in a collaborative environment like Wikidata, it’s also very important to have periodic monitoring of statements, translations and references needed in order to obtain accurate information from these assistants. Another major direction is to have a greater participation from public-private partnerships

including tourism agencies. For this purpose, generic and simpler tools to monitor urban data information as well as to understand and detect possible vandalism events need to be explored.

8 https://doi.org/10.5281/zenodo.1252427

Table 1. Properties used on Wikidata for urban information

Datatype	Properties
CommonsMedia	image (P18), flag image (P41), coat of arms image (P94), seal image (P158), locator map image (P242), pronunciation audio (P443), page banner (P948), spoken text audio (P989), collage image (P2716) nighttime view (P3451)
ExternalId	ISNI (P213), VIAF ID (P214), GND ID (P227), Library of Congress authority ID (P244), Biblioth`eque nationale de France ID (P268), SUDOC authorities ID (P269), ISO 3166-2 code (P300), OSM relation ID (P402), Freebase ID (P646), NKCR AUT ID (P691), FIPS 10-4 (countries and regions) (P901), SELIBR ID (P906), Biblioteca Nacional de Espan˜a ID (P950), MusicBrainz area ID (P982), DMOZ ID (P998), AAT ID (P1014), BAV ID (P1017), US National Archives Identifier (P1225), WOEID (P1281), Gran Enciclop`edia Catalana ID (P1296), Encyclopædia Britannica Online ID (P1417), GeoN-ames ID (P1566), TGN ID (P1667), Global Anabaptist Men-nonite Encyclopedia Online ID (P1842), Facebook Places ID (P1997), GRID ID (P2427), Great Russian Encyclopedia On-line ID (P2924), Encyclopædia Universalis ID (P3219), NE.se ID (P3222), Quora topic ID (P3417), archINFORM location ID (P5573), Petit Fut´e site ID (P5760), Dizionario di Storia Trec-cani ID (P6404), Who's on First ID (P6766), ROR ID (P6782), DANE code (P7325)
GlobeCoordinate	coordinate location (P625)
Monolingualtext	official name (P1448), demonym (P1549), native label (P1705), short name (P1813)
Quantity	population (P1082), length (P2043), elevation above sea level (P2044), area (P2046), width (P2049)
String	postal code (P281), Commons category (P373), licence plate code (P395), local dialing code (P473), Commons gallery (P935), Commons maps category (P3722)
Time	inception (P571)
Url	official website (P856)
WikibaseItem	head of government (P6), country (P17), instance of (P31), official language (P37), shares border with (P47), founded by (P112), located in the administrative territorial entity (P131), contains administrative territorial entity (P150), flag (P163), twinned administrative body (P190), legislative body (P194), located in or next to body of water (P206), executive body (P208), coat of arms (P237), part of (P361), located in time zone (P421), said to be the same as (P460), opposite of (P461), topic's main category (P910), topic's main Wikimedia portal (P1151), office held by head of government (P1313), described by source (P1343), capital of (P1376), present in work (P1441), category for people born here (P1464), category for people who died here (P1465), category of people buried here (P1791), cat-egory of associated people (P1792), owner of (P1830), different from (P1889), history of topic (P2184), on focus list of Wikime-dia project (P5008)

Table 2. Translation of labels, description and aliases on Wikidata for urban informa-tion of selected Colombian cities

Measure	Labels	Description	Alias
count	40	40	40.000000
mean	54.70000	11.500000	9.375000
std	38.87996	5.223222	11.331118
minimum	1	0	1
25%	21.75	9.75	2.00
50%	56.50	10.50	4.50
75%	69.25	12.00	12.000
maximum	223	32	54

Table 3. Statements, References, Inbound Links, Wikimedia articles on Wikidata for urban information of selected Colombian cities

Measure	Statements	References	Inbound Links	Wikimedia articles
count	40.000000	40.000000	40.000000	40.000000
mean	154.775000	43.450000	480.625000	45.375000
std	102.644142	44.839857	1195.322158	39.173963
min	13.000000	1.000000	1.000000	1.000000
25%	85.750000	14.750000	49.500000	20.000000
50%	145.500000	18.000000	103.000000	38.500000
75%	176.750000	73.000000	332.000000	51.750000
max	584.000000	188.000000	7048.000000	235.000000

BIBLIOGRAPHY

Li Ding, Vassilios Peristeras, and Michael Hausenblas. Linked Open Govern-ment Data [Guest editors' introduction]. IEEE Intelligent Systems, 27(3):11–15, May 2012.

Dustin Lange, Christoph B¨ohm, and Felix Naumann. Extracting structured information from Wikipedia articles to populate infoboxes. In Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10, page 1661, Toronto, ON, Canada, 2010. ACM Press.

Marshall McLuhan and Bruce R. Powers. The global village: transformations in world life and media in the 21th century. Communication and society. Oxford Univ. Press, New York, 1992. OCLC: 845334715.

Santiago M. Mola-Velasco. Wikipedia vandalism detection. In Proceedings of the 20th international conference companion on World wide web - WWW '11, page 391, Hyderabad, India, 2011. ACM Press.

Joachim Neubert. Wikidata as a linking hub for knowledge organization sys-tems? integrating an authority mapping into wikidata and learning lessons for KOS mappings. In Proceedings of the 17th European Networked Knowl-edge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), Thessaloniki, Greece, September 21st, 2017., pages 14–25, 2017.

John Samuel. Analyzing and visualizing translation patterns of wikidata properties. In Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings, pages 128–134,
2018.

Thomas Pellissier Tanon and Lucie-Aim´ee Kaffee. Property label stability in wikidata: Evolution and convergence of schemas in collaborative knowledge bases. In Companion of the The Web Conference 2018 on The Web Confer-ence 2018, WWW 2018, Lyon , France, April 23-27, 2018, pages 1801–1803,
2018.

Denny Vrandecic and Markus Kr¨otzsch. Wikidata: a free collaborative knowl-edgebase. Commun. ACM, 57(10):78–85, 2014.

Fei Wu and Daniel S. Weld. Automatically refining the wikipedia infobox ontology. In Proceeding of the 17th international conference on World Wide Web - WWW '08, page 635, Beijing, China, 2008. ACM Press.

PEDESTRIAN BEHAVIOUR MODELING AND SIMULATION FROM REAL TIME DATA INFORMATION

MICHAEL PUENTES2[0000–0002–1802–839X], DIANA NOVOA2, JOHN M. DELGADO NIVIA2, CARLOS J. BARRIOS HERNNDEZ1, OSCAR CARRILLO3, AND FRDRIC LE MOUEL3

1 UNIVERSIDAD INDUSTRIAL DE SANTANDER CBARRIOS@UIS.EDU.CO

2 UNIDADES TECNOLGICAS DE SANTANDER TERRITORIOINTELIGENTE@UTS.EDU.CO

3 INSA - LYON OSCAR.CARRILLO@INSA-LYON.FR,FREDERIC.LE-MOUEL@INSA-LYON.FR

ABSTRACT.

Accidents of pedestrians sometimes take lives, in Bucara-manga since 2012 pedestrian died by accidents are 179, and 2873 hurt, In a city as Bucaramanga, this means each day at least one pedestrian is involved in a accident. Therefore is necessary to know the causes of accidents in the way to decrease the accidents. One of many reasons to know the causes is with system dynamics, simulating the events of the Pedestrian behavior when accidents occur in risen cities. The implemen-tation simulation joint with technology and research looking for saving lives, reducing the accidental rate, and to implementing or suggesting new policies from the government. This project is looking for the imple-mentation of technology in video records and Deep Learning analysis for the service of the citizens, where a simulation model will be revealing the main variables which intervene in the pedestrian's behavior. As initials results, shows the methodology here implemented, can reach data which was insufficient before thanks to the cameras and software of objects de-tection, those are the data input for the simulation model, which after to implement a change in a particular spot of Bucaramanga is possible decrease the accident rate in 80% where pedestrians could be involved.

Keywords: Pedestrian Behavior · traffic violation · Dynamic System

1. INTRODUCTION

The approach in this project is looking for related research about pedestrian behavior in an urban area, with the aims to reduce accidents in a city. Yang et al [1] made an important contribution, separating two types of pedestrians, the obey law ones or the opportunistic ones, this is an important criterion be-cause is the perception of the pedestrians a city as Bucaramanga in Colombia, have problems with the obey of the traffic authorities as traffic lights.

This is just an assumption of the idea, thus, they made a questionnaire evaluating the behavior of the people of chine related to the pedestrian cross path. In this ques-tionnaire, some variables are important in the construction of the model for the micro-simulation, as age, and gender among others. Another related work which evaluates the behavior of pedestrians splited by gender and age, this research work made it by

Chen Chai et al [2], both variables have extra information of the model, which are children and gender from Fuzzy logic-based observation. From this work, born the question of the variables who affect the behavior of the pedestrians. Aaron et al [3] is a work which for the comparison of the reality which is possible to count the variables which are related to the environment, knowing the reality always will be change.

This work determines the variables of a micro-simulation which will be part of the causal model for the refined reality. moreover, Camara et al [4], implemented a decision tree to determine the pedes-trians' vehicle interaction, this is an important implementation, looking for the designs of new policies for pedestrians in Bucaramanga - Colombia, looking for less critical accidents where pedestrians are involved (see fig. 1), Holland et al [5], see the gender as an important factor in the behavior of a pedestrian in the decision of crossing a pedestrian crossing in a simulation study. In others works the micro-simulation are made with a software it can know the pedestrians event drivers behavior as individuals [6, 7], is the PVT software with the modules of VISSIM and VISWALK.4

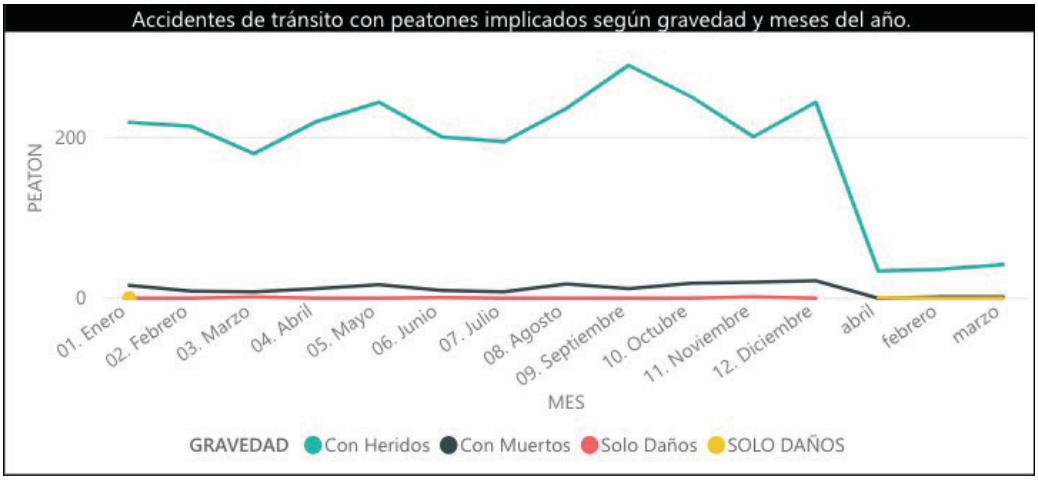


Fig. 1. Pedenstrians' accidents from 2012 in Bucaramanga.
(from: http://observatorio.bucaramanga.gov.co/index.php/informacion-publica/)

2. PREVIOUS WORKS

The aim of the research will be to identify the variables which intervene in the jaywalking of pedestrians or external causes that produce accidents in Bu-caramanga city.

To find this causes it will find different scenarios with real-life information and people perception, and hereafter to validate these variables with a micro-simulation model, which will be refined the reality meanwhile is comparing with video-recordings in a spiral process of refined simulation by Jordan (see fig 2),

Where the validation is repeatedly processed in as endless task, because citizens behavior pattern is not trivial, nevertheless, with a near prediction of the pedestrian behavior it is possible, for to implement new policies and campaigns in the city to improve the social behavior of the citizens. As further work, it will determine new variables which are possible be in the model across the citizens perception, and this will be compared with the made model [8].

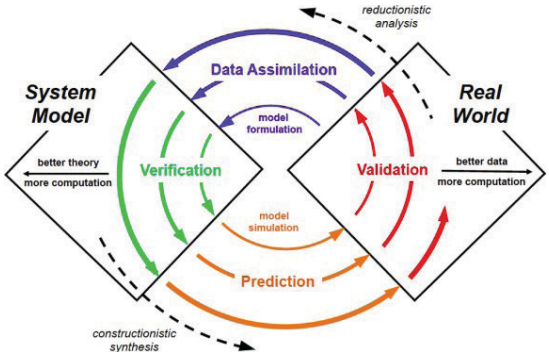


Fig. 2. Inference Spiral of System Science (Jordan 2015)

In the interest to minimize the accidents who involved citizens and vehicles, it is necessary to find the reasons which affect the decision of pedestrians oppor-tunistic ones, or jaywalking, when the traffic light allows the pass of the vehicles or scenarios where do not exists a traffic lights, but the pedestrians must have priority in the way, and the vehicle will stop in that event.

A previous work to starts is the empirical analysis of Sanghamitra et al [9], who list the crossing decisions of pedestrians in a cross side of the street based on time gap until the arrival of the next car, where not all the variables could detect it with a tool of the research, but it is possible to recognize a similar behavior in the pedestrians. Another well-know variable which takes place in the pedestrian behavior is the ”social force”, it occurs when the pedestrians are guided by another citizen, with-out knowing if the principal citizen decision is correct, but at least have a better vision of the path.

Different models for pedestrians can be found Magnetic force, Social force and Benefit-Cost Cellular in the literature. those areas searchable where Teknomo does a review of microscopic simulation of pedestrian, detailing every pedestrian as individual [10], in this work, different variables are necessary for a mathematical model, but different causes could be part of the event of a pedestrian accident. In a simulation model the next causes could be considered in a simulation model:

- Time gap between car and pedestrian [1, 9]
- Social Force [10, 11]
- Environment (Weather, pollution, noise) [12]
- Vehicle Factors [12]
- Human Factors (Driver Skills, Fatigue, Alcohol, drugs, failed looks properly) [12, 13]
- Road Conditions (Corner, straight, wet, dry) [14]

This kind of causes must be identified from different scenarios where the pedestrian can have a particular behavior. Pau et al [15], select the scenarios from a different time of the day, which there is a big quantity of pedestrian in the streets (peak hours) between other time when pedestrians quantity is low. Rasouli et al [16], determines the people who cross-traffic lines in the street and who made a signal with the hands, indicating a petition of the stop to the driver, by the methodology of this research it is necessary looking for different environments, night, day, rain, snow.

This work will be no related with the implementation of objects in the image, but, if the action of the pedestrian change in the darkness, or persist, is an environment which is necessary to include in the

methodology. In this case, the scenarios in the same place will be during day comparing with the night in certainly hour when pedestrians are crowded, and when not. Kouabenan et al [14], they analyze 55 reports of pedestrians' accidents, randomly selected from a police report in the Ivory Coast, in this research, they analyze the characteristics and circumstances of the accidents.
Thereby, there those previous works, conclude the research, it suggests new solutions in the city, which simplify in new public policies, campaigns, or im-provement in a particular spot of the city of the research, this shows them in the next list:

- Accident prevention campaigns [14, 17]
- Road safety policies [12, 18]
- Improve lighting conditions [18, 19]
- Vehicle conditions campaigns [18]

3. METHODOOGY

This research work consists in to create a micro-simulation model with the factors which intervene in the citizen behavior (pedestrian or driver) and to refine the simulation with the real-life through some cameras installed in different spots in 5 cities of Colombia, initially: Bucaramanga. This work uses the paradigm of system dynamics; the analysis consists of the following 4 steps:

1. Reviewing the influences on the phenomenon witnessed
2. Modify the simulation model
3. Evaluate the equations created from here
4. The behavior that the observed phenomenon will have, compared with the model created

This process previously described, will be a cycle that will be perfected when the model can add more variables according to the observations of the phe-nomenon, see Fig 3. If in the simulation it is possible to find the causes who

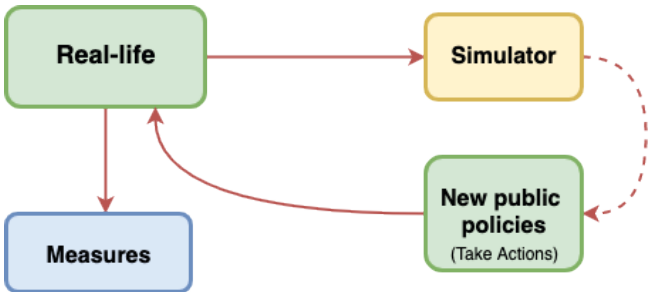


Fig. 3. General Methodology Implemented

have more impact on the pedestrians' accidents in the city, then that causes will be transformed in new public policies to alter the reality and decrease the pedestrians' accidents.

Across a project in the metropolitan area of Bucaramanga by Government of Santander, we have available 900 cameras in 5 cities of Santander-Colombia, for the analysis of the citizens, with the aim of improving civics in the city. Different aims are proposals through those cameras, mobility, public spaces and harmony. With the aim of further a better mobility and to reduce the rate of accidents in the city, this research focuses on pedestrians, thus it decides to select from those cameras the spots were to occur more accidents or an

spot where have different conditions. according to the fig 4, the selected camera has the image show it in the fig 5. This particular camera has no traffic lights and is in the zone which collects different variables according to the previous work.

Also, it is the second zone with more accidents and has a mixture of architecture, with means, one church, 2 universities, a park, and near, is full of particular homes.

The next step is the scenarios where the video recording of those cam-eras, it proposes two main scenarios: a main hour in the city and illumination (day/night), the combination of these probabilities is four videos per day for comparing the behavior. Then it selects 10 days to do the comparison that rep-resents 40 videos (1 hour each), in each video-recording, it checks variables to identify the representation of the simulation model according to this methodol-ogy, the theoric variables will be initial meanwhile it refines with the real-life in the observations of the video. with all the videos we will have an expectation to found a behavior pattern with the aim to feed the simulation with real and accurate information.

Initially, the simulated model starts from the previous researches carried out to the evaluated phenomenon, bearing in mind that all the models can vary, because the model is created for the particular case of the city of Bucaramanga. From the real-life causes, it can see the measurable causes because of the available cameras of the city. It has tools for the measurable causes, which are video-recordings and software to find objects trow deep learning, this software will rush it the process to find pedestrians in hours of video (Briefcam5). Then, with the video-recording and the software, it is possible to obtain the data which is detailed in table 1. this is a piece of quantifiable information, but it is also necessary to have extra information from particular behavior from the video, and the perception of the people who live in Bucaramanga.

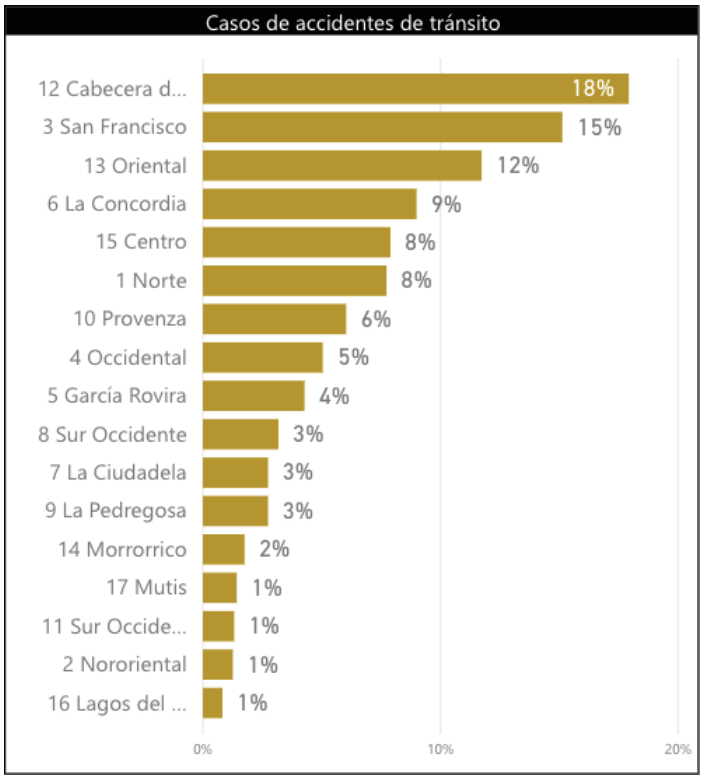


Fig. 4. most accidental areas in Bucaramanga
(from: http://observatorio.bucaramanga.gov.co/index.php/informacion-publica/)

3.1 MODEL AND SIMULATION

According to the previous work reviewed in chapter 2, it gathers some causes who directly entail in the pedestrians' accident, who are the aim of this research. otherwise the research will estimate it, if is necessary a micro-simulation [20], looking the pedestrian as an individual, or to identify the general causes seen in the simulation observed, and measurable in the cameras installed for this project (see Fig. 6). In previous researches, many simulations were implemented [8, 12, 16–18] and the solution implemented are new policies, which its solution have a senoidal behavior according to Mendez[17], thereby, is necessary to implement a micro-simulation to see the pedestrian's causes of accident as an individual [1, 10] to determine the causes as individual trying to minimize the accident rate in the city.
5 https://www.briefcam.com/

\\

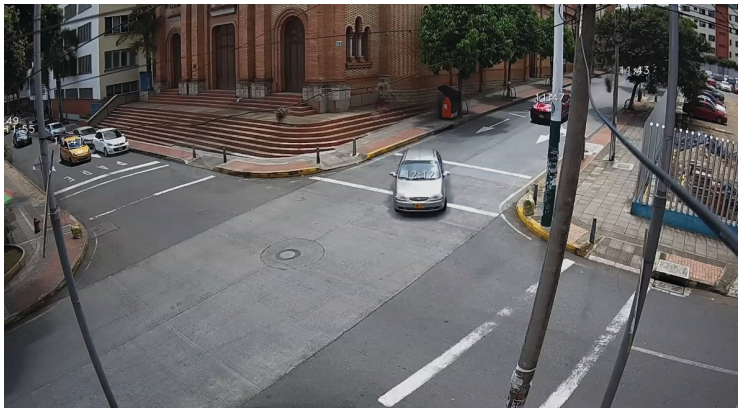


Fig. 5. Vision angle of video camera selected

Variable	Type	Observation
Date of Video	Date	
Range hour	e.g: 21:00 - 22:00 hrs	
Pedestrians who cross the street	numeric	categorized by gender
Not safe events	numeric	those are events which involves an accident
pedestrians against the law	numeric	
pedestrians who cross by the zebra	numeric	
Pedestrians velocity aver-age	numeric	

Table 1. variable of the pedestrian

In the case of a micro-simulation, causes are directly related to variables measurables and which are possible to get it in a video recording, then, from the initial causes evaluated, this work will assess the human factors causes in the micro-simulation for pedestrians as individuals. The numeric data information is not enough, that's why we use the perception of the people to know extra information which is not visible in the videos.

In the accidents some pedestrians endangering their own lives, they interfere with traffic in some way, thus, the motivation is looking for safe crossing. Then through the Viswalk simulation tool6, it will find the causes of accidents in a

6 https://www.ptvgroup.com

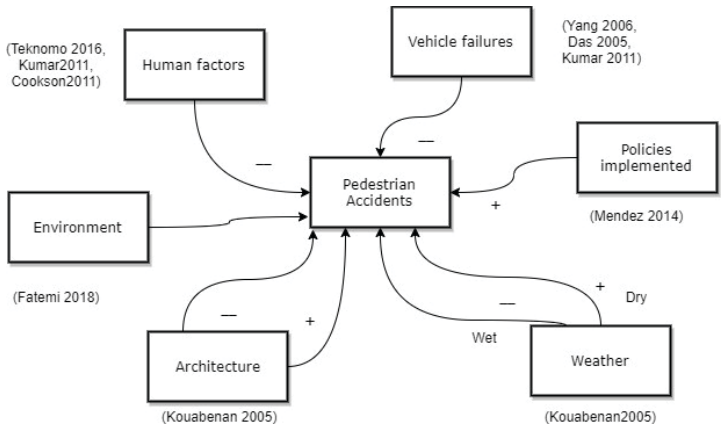


Fig. 6. Causal Diagram for pedestrian accidents

micro-simulation of pedestrians in a specific sector of Bucaramanga-Colombia, using the same methodology represented in Fig 3. First, it will analyze priority, which in Colombia the priority has the cars instead of the pedestrians.

Fig. 7. viswalk micro-simulation in San Francisco neighborhood

To know the causes who a jaywalker could have, we ask to students, and community in general, about how good pedestrians are. the result show in the 2 and 3. This information was used to feed the micro-simulation, for example, and one of the more exciting information, just 69% of people think, the pedestrian is not respected, even in a zebra, the car or motorbike has the priority. This is an important information, thereby, when the simulation have this rule of priority, the accidents in the micro-simulation starts to appear.

Question	always	often	sometimes	rarely	never
Do you walk on the ze-					
bra crossing when you cross	42.9%	47.6%	9.5%	0.0%	0.0%
road?					
Do you look at the state of					
traffic lights when you cross	95.2%	4.8%	0.0%	0.0%	0.0%
road? Answers					

Table 2. Pedestrians' behavior at intersections

Question	always	often	sometimes	rarely	never
In general conditions	0.0%	9.5%	23.8%	33.3%	33.3%
In a hurry	14.3%	23.8%	38.1%	19%	4.8%
Long duration of red light	14.3%	14.3%	9.5%	33.3%	28.6%
Presence of other pedestri-					
ans who violate traffic sig-	9.5%	9.5%	9.5%	19%	52.4%
nal					
Low traffic volume	33.3%	38.1%	14.3%	4.8%	9.5%
High traffic volume	19%	9.5%	0.0%	23.8%	47.6%
Policeman is on duty at the	23.8%	4.8%	14.3%	9.5%	47.6%
intersection					

Table 3. Probabilities of pedestrians' signal non-compliance under specific situations

The quantifiable information from the video recording, from the spot of the city, is showed in the table 4. With the information about the perception of the people and the numeric information, the microsimulation starts to have information. In order to reach any kind of change behavior, we include a road speed reducer in one of the vehicular streets, where they have to decrease the speed to 2km/hr. This small but significant change in the ”reality” according to the simulation is possible to save more than 80% of the people related in an accident in that particular spot with the associate information.

Variable	Type	Observation
Date of Video Range hour	22/06/2019 e.g: 9:00 - 10:00 hrs	
Pedestrians who cross the street	316	70% man and 30% woman
Not safe events	0	not registered in the video
pedestrians against the law	2	
Vehicles in the video	668	
cars and motorbikes vehi-cles		
velocity average	60km/hr	

Table 4. The quantifiable information from the video recording

4. CONCLUSIONS

The system model created, will help to determine the fittest variables which are more important in the act of unsafe crossing of the pedestrians in Bucaramanga, moreover, the system created will be transferred to another place with a similar environment where is possible to have the same scenarios to compare the pedes-trians' conduct, looking to find a pattern behavior. The implementation with a small change in architecture it means a significant number of saving people's lives. Then the methodology proposed is a good first step to use Infrastructure (cameras) and information (video-recording) in order to build a smart city in Bucaramanga.

5. FUTURE WORK

With the methodology implemented, it is possible to analyze different causes seen in this research, associated with weather, architecture, people behavior, vehicle conditions, even policies implemented.

6. ACKNOWLEDGE

Thank you to the Government of Santander with the project 879/2017 between the government and the Unidades Tecnologicas de Santander. Thank you SC3-UIS, Colifri, CitiLab-INSA and CATAI where was presented this project.

REFERENCES

J. Yang, W. Deng, J. Wang, Q. Li, and Z. Wang, "Modeling pedestrians' road cross-ing behavior in traffic system micro-simulation in China," Transportation Research Part A: Policy and Practice, vol. 40, no. 3, pp. 280–290, 2006.

C. Chai, X. Shi, Y. D. Wong, M. J. Er, and E. T. M. Gwee, "Fuzzy logic-based observation and evaluation of pedestrians' behavioral patterns by age and gender," Transportation Research Part F: Traffic Psychology and Behaviour, vol. 40, pp. 104–118, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.trf.2016.04.004

M. A. Aaro´n, C. A. Go´mez, J. Fontalvo, and A. J. Go´mez, “Ana´lisis de la Movilidad Vehicular en el Departamento de La Guajira usando Simulacio´n. El Caso de Riohacha y Maicao,” Informacio´n tecnolo´gica, vol. 30, no. 1, pp. 321–332, 2019.
F. Camara, O. Giles, R. Madigan, M. Rothmuller, P. H. Rasmussen, S. A.
Vendelbo-Larsen, G. Markkula, Y. M. Lee, L. Garach, N. Merat, and C. W. Fox, “Predicting pedestrian road-crossing assertiveness for autonomous vehicle control,” IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, vol. 2018-Novem, pp. 2098–2103, 2018.
C. Holland and R. Hill, “Gender differences in factors predicting unsafe crossing
decisions in adult pedestrians across the lifespan: A simulation study,” Accident Analysis and Prevention, vol. 42, no. 4, pp. 1097–1106, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.aap.2009.12.023
J. Oskarbski and L. Gumin´ska, “The application of microscopic models in the study
of pedestrian traffic,” MATEC Web of Conferences, vol. 231, pp. 1–7, 2018.
T. Campisi, G. Tesoriere, and A. Canale, “The pedestrian micro-simulation applied to the river Neretva: The case study of the Mostar ”old bridge”,” AIP Conference Proceedings, vol. 2040, no. November, 2018.
V. Cantillo, J. Arellana, and M. Rolong, “Modelling pedestrian crossing behaviour
in urban roads: A latent variable approach,” Transportation Research Part F: Traffic Psychology and Behaviour, vol. 32, pp. 56–67, 7 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1369847815000716
S. Das, C. F. Manski, and M. D. Manuszak, “Walk or wait? An empirical analysis
of street crossing decisions,” Journal of Applied Econometrics, vol. 20, no. 4, pp. 529–548, 2005.
K. Teknomo, Y. Takeyama, and H. Inamura, “Review on Microscopic
Pedestrian Simulation Model,” no. March, pp. 1–2, 2016. [Online]. Available: http://arxiv.org/abs/1609.01808
D. Helbing and P. Molna´r, “Social force model for pedestrian dynamics,” Physical
Review E, vol. 51, no. 5, pp. 4282–4286, 1995.
N. . Kumar and G.Umadevi, “Application of System Dynamic Simulation Modeling in Road Safety,” Trasnportation Research Board, 2011.
R. Cookson, D. Richards, and R. Cuerden, The characteristics of pedestrian road
traffic accidents and the resulting injuries, 2011.
D. R. Kouabenan and J.-M. Guyot, “Study of the causes of pedestrian accidents by severity,” Journal of Psychology in Africa, vol. 14, no. 2, 2005.
G. Pau, T. Campisi, A. Canale, A. Severino, M. Collotta, and G. Tesoriere, “Smart
pedestrian crossing management at traffic light junctions through a fuzzy-based approach,” Future Internet, vol. 10, no. 2, 2018.
A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Understanding Pedes-
trian Behavior in Complex Traffic Scenes,” IEEE Transactions on Intel-ligent Vehicles, vol. 3, no. 1, pp. 61–70, 3 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8241847/
G. M´endez-Giraldo and L. A´lvarez-Pomar, “Dynamic model to analyze pedestrian traffic policies in Bogota,” Dyna, vol. 81, no. 186, p. 276, 2014.
G. F. Ulfarsson, S. Kim, and K. M. Booth, “Analyzing fault in pedestrian-motor vehicle crashes in North Carolina,” Accident Analysis and Prevention, vol. 42, no. 6, pp. 1805–1813, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.aap.2010.05.001
G. Zhang, K. K. Yau, and X. Zhang, “Analyzing fault and severity in pedestrian-motor vehicle accidents in China,” Accident Analysis and Prevention, vol. 73, pp. 141–150, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.aap.2014.08.018
L. A´lvarez-Pomar, “Modelo de inteligencia colectiva de los sistemas peatonales,”
Universidad Distrital Francisco Jose de Caldas, vol. 23, no. 45, pp. 5–24, 2016.

BIM-BASED MIXED REALITY ENVIRONMENTS TO IMPROVE AEC TASK PERFORMANCE

PIERRE RAIMBAUD1,2[0000–0002–5584–8100], FREDERIC MERIENNE1[0000–0003–4466–4776], PABLO FIGUEROA2[0000–0001–5412–8630], FLORENCE DANGLADE1, RUDING LOU1, AND JOSE TIBERIO HERNANDEZ2[0000–0002–5035–4363]

1 LISPEN, ARTS ET METIERS, INSTITUT IMAGE, CHALON-SUR-SAONE, FRANCE PIERRE.RAIMBAUD@ENSAM.EU, FREDERIC.MERIENNE@ENSAM.EU,

RUDING.LOU@ENSAM.EU,FLORENCE.DANGLADE@ENSAM.EU

2 SYSTEMS AND COMPUTING ENGINEERING, IMAGINE GROUP, UNIVERSIDAD DE LOS ANDES,

BOGOTA D.C., COLOMBIA P.RAIMBAUD@UNIANDES.EDU.CO, PFIGUERO@UNIANDES.EDU.CO, JHERNAND@UNIANDES.EDU.CO

ABSTRACT.

The Building Information Modelling (BIM) currently con-tributes to deeply modify the Architecture, Construction and Engineer-ing (AEC) industry by improving data management, task planning, and architecture design etc. Nevertheless, other technologies have also joined this revolution, with the aim of allowing experts to perform better their tasks with them than with only the BIM, particularly mixed reality (MR).

However, MR applications can take very diverse forms, because of the multiple design choice possibilities: multiple data sources from the BIM (3D model, worksite monitoring, simulations. . .), multiple possi-bilities of visualisations in MR (visual effects, 4D. . .) and multiple MR interactions (move, write, say, grasp. . .). Behind MR application design choices, there is a task for which the application has been created. Yet, having BIM-based MR environments that really allow to respond to the original need and that improve task performance is a current difficulty. In this paper, we present our proposal of a methodology for going from BIM to BIM-based mixed reality environments.

Our inputs are the AEC tasks which are likely to benefit from being performed in a mixed reality environment, their performance measures (efficiency and effectiveness), and BIM data.

Our target is to provide BIM-based mixed reality en-vironments that support specific AEC tasks, and to prove thanks to appropriate indicators that the task performance has improved in MR compared to traditional methods. Thus, we present here the results from our first case studies and their impact on the methodology evolution. Finally, our ongoing and future works are discussed in the last sections.

Keywords: BIM · mixed reality · design choice · task performance

1. INTRODUCTION
1.1 CONTEXT

The Architecture, Construction and Engineering (AEC) industry is a sector which has changed in the last decades, particularly with the Building Information

Modelling (BIM), a new concept or methodology. This relies on centralising all the construction project data, and updating them during the whole life-cycle of the building. Data can be interdisciplinary models (2D and 3D), costs, supplier contracts, and a building life-cycle is composed of five main phases [1]: the preconception phase, the design phase, the construction phase, the operation, and the maintenance.

For all phases, the BIM can help, and for that, many BIM tools are used, each one having its own purpose, from energy analysis to health and safety management. Nonetheless, other technologies can also help and be complementary with BIM technology, such as machine learning tools [2], or sensor technologies and the Internet of Things [3]. In our current research project, we focus on mixed reality (MR), which allows to create virtual environments.

These can be made from data coming from reality, from virtual modelling, or from both. This is usually called the virtuality-reality continuum [4] ; the term mixed reality encompasses both the concept of augmented reality [4] and virtual reality [4]. Thus, in a MR application, data can come from multiple sources ; in the AEC field, it can be 3D models from a BIM authoring software [5] (virtual), or videos from drones (real). Additionally, the visualisations chosen to show these data, and the interactions provided to the users can also be multiple. Thus, design choices are bountiful when creating MR applications for AEC purposes.

12 PROBLEM

The BIM methodology and the BIM tools allow the AEC industry to improve the construction process, regardless of the phase, the involved stakeholders or the performed tasks [6][7]. However, for some tasks, human expertise is and will still be mandatory, no matter the help that the BIM tools can supply ; in these cases, BIM tools and data could be used either as inputs, or as automated ap-proximated results, to facilitate decision-making.

Additionally, as shown in the literature, other technologies could be linked to BIM to enhance human expertise, such as Geographic Information Systems (GIS) [8], or mixed reality [9][10]. In these two last studies, the authors justified the added-value of MR by measur-ing the performance of the experts with their application when performing and comparing it with the results using traditional methods. Here the measures are duration, number of correct answers [9] and time, errors, interaction accuracy, subjective questions about collaboration effectiveness [10].

Thus, according to this current state, some may wonder how MR should be used in a BIM context and how to evaluate its added-value. Based on these two questions, we seek for a methodology for designing MR applications in an ”AEC-BIM context”, and to ensure valuable implementations. For that, we also ask for using measures to both estimate the added-value of mixed reality (in terms of efficiency and effectiveness) and also to improve continuously this methodology.

2. METHODOLOGY

In this research project, we propose a methodology for obtaining MR design choices through an analysis of the AEC tasks for which a BIM-based MR application is developed. Then, we propose to apply it on case studies, resulting in building MR applications prototypes.

Finally, we propose to conduct exper-iments in these MR environments measuring task performance to evaluate and compare them to traditional tools, in terms of efficiency and effectiveness. With these applications, we aim to help to tackle current gaps in civil engineering.

In this paper, we want first to present our methodology. As our BIM-based MR applications aim allow one or several kinds of stakeholders of an AEC project to better complete a specific task, our methodology relies on different taxonomies about tasks present in the literature: taxonomies of AEC tasks [11], taxonomies of generic task [11][12], and taxonomies of tasks in mixed reality [12][13][15][16]]. Figure1 presents our methodology steps: first, our idea is to go from a field-specific naming (sometimes very close to the AEC jargon), to a generic naming, thanks to the abstraction and decomposition of AEC tasks into generic simplest subtasks.

Then, these subtasks (WHY) can drive design choices for the applica-tions, helping to make decisions about 1) the content/data (WHAT), and 2) the idioms (HOW) - a set of idioms can be defined as distinct approaches for creating or manipulating visual representations, i.e. the interactions (how to perform) and the visualisation (how to represent). Finally, MR applications are built based on these WHAT and HOW, and we measure the quality of the applications, and of the methodology itself, to repeat and improve this methodology.

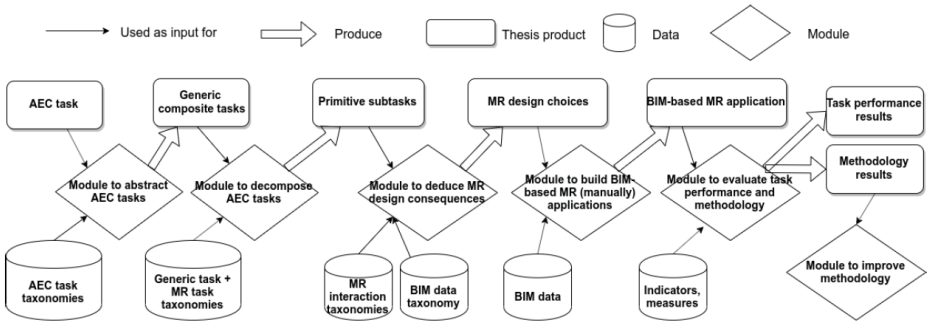


Fig. 1. Our methodology for building effective and efficient BIM-based mixed reality environments, according to the AEC tasks

3. RESULTS FROM PRELIMINARY CASE STUDIES

As preliminary work, we built and evaluated MR application prototypes, for three different case studies. Note that these studies led to three scientific pub-lications [19][20][21]. Then, usability and efficiency measures applied in each experiment helped us to build, correct, and formalise incrementally our method-ology, resulting in what is explained in this paper (Figure 1). In this section, we present these cases studies and their MR applications.

3.1 ARCHITECTURE DESIGN REVIEW OF A NURSERY

Our first case study (Figure 2) was about a nursery construction project, where the architect and future final users wanted to review the architectural design according to the future usages (WHY). If we would have used task abstraction and decomposition, we would have had as subtasks: navigate in architectural model, check light exposition, confirm design, and check building usability for future operation. In this study [19], we wrote a previous version of our method-ology, wondering which BIM data would be useful, how it should be shown, and how to perform in this environment. The data (WHAT) used were sun simula-tions and the architectural BIM model, interactions (HOW) were navigation by teleportation, and visualisation was realistic (shadows). However, some interac-tions were missing (annotate), and measures were not taken neither to compare performance (time, number of annotations), nor for improving our methodology.

Fig. 2. Architecture design of the nursery in our virtual reality application [19]

3.2 CONSTRUCTION SUPERVISION FROM BIM MODELS AND DRONE VIDEOS

Our second case study [20] focused on improving construction supervision, which is necessary to ensure planning enforcement, by allowing off-site supervision and comparisons between as-built (from drone videos) and as-planned design (from BIM models). Thanks to superimposition in MR (Figure 3, b) and c)), inspectors could estimate differences to make their report and enrich the BIM model.

Fig. 3. From left to right: a) video of the building from a drone, b) MR visualisation: superimposition of the video and the BIM model c) MR interaction: annotations [20]

We understood with this study that design choices were driven by the AEC task: real and virtual data were necessary (WHAT), navigation by points of interest was very useful to focus only on the new elements at T time (HOW-interactions), and superimposition with transparency was mandatory to see dif-ferences (HOW-visualisation). In this study, measures for evaluating the method-ology were taken (positive feedback) but comparative measures missed.

4 ONGOING AND FUTURE WORK

Thanks to our methodology, we have seen that we can more easily build MR prototypes ; thus, in this research project, we propose to focus on two chal-lenges: 1) the design clash analysis and 2) the dynamic hazard identification, both identified as gaps in the civil engineering literature [17][18].

4.1 ONGOING WORK: BIM DESIGN COORDINATION, CLASH ANALYSIS

BIM models coordination is mandatory since experts usually made their design independently and then fusion them, resulting in an interdisciplinary model. Thus, design clashes are common, but they can be detected automatically by current BIM tools. However, clash resolution still requires human expertise. In this third case study [21], by applying our methodology, it is possible to receive design choices, such as discrimination colours between discipline (HOW – vi-sualisation), free movements and annotation positioning (HOW – interactions). Figure 4 shows our MR application following them.

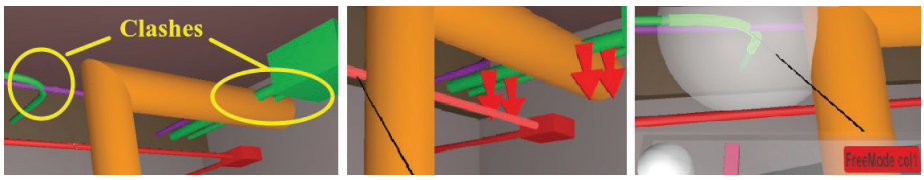


Fig. 4. From left to right: clashes examples; annotations in MR: narrows, spheres [21]

Our priority now is to implement ways to measure (and compare) both task performance and methodology accuracy. A preliminary experiment tended to indicate that experts solved the design clashes faster (measure: time) and better (measure: grade solution) with our MR application rather than using current computer screen tools.

Nonetheless, the application must still be improved and the experiment repeated with more experts to get more conclusive results.

4.2 FUTURE WORK: HAZARD IDENTIFICATION

Our future case study is about the hazard identification task performed by safety inspectors (specialised civil engineers).

If we apply abstraction and decomposi-tion as explained in our methodology, this task is composed by substasks such as observing the hazardous zones, analysing the hazards, and estimating their probability, severity and size... Thus, task performance could be measured us-ing indicators such as the time spent to identify the hazards, and the number of (correct) hazards detected, and compare with the results with current tools.

According to the substasks, two or three inputs for the application (WHAT) are required: the design model of the construction, the 4D BIM simulation (over time), and if possible automated hazard partial detection. Expected results are that, with our tool, it will take less time and more hazards will be identified.

CONCLUSION

The BIM allows to improve the construction process through data centralisation and computational power for simulations. Mixed reality can also help by im-proving task performance; nonetheless, a MR tool is usually created to perform one task. Thus, AEC tasks must be understood for a great integration in a MR environment, and this environment must be created accordingly.

That is why we proposed a methodology where AEC tasks are abstracted and decomposed into subtasks, which are implementable in MR, and which induce MR design choices. Given this method, we proposed to apply it on two different cases to try to tackle the challenges they presented: design clash analysis and hazard iden-tification. As future line of research work, we propose to build these two MR applications, according to the design choices coming from our methodology, and to evaluate and compare task performance using them or traditional tools.

REFERENCES

X.Xu, L.Ma, L.Ding, A framework for BIM-enabled life-cycle information man-agement of construction project (2014) International Journal of Advanced Robotic Systems, 11 (1), art. no. A126, https://doi.org/10.5772/58445

A.Braun, A.Borrmann, Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning, Au-tomation in Construction, vol. 106, (2019), 102879, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2019.102879

S.Tang, D.R. Shelden, C.M. Eastman, P.Pishdad-Bozorgi, X.Gao, A review of build-ing information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends, Automation in Construction, vol. 101, (2019), pp. 127-139, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2019.01.020

P. Milgram, H. Takemura, A. Utsumi, F. Kishino, Augmented Reality: A class of displays on the reality-virtuality continuum, (1994) Proceedings of Telemanipulator and Telepresence Technologies, vol. 2351, https://doi.org/10.1117/12.197321

https://www.accasoftware.com/en/bim-authoring Last accessed on 20/10/19

S.Azhar, Building information modeling (BIM): Trends, benefits, risks, and chal-lenges for the AEC industry, (2011) Leadership and Management in Engineering, 11 (3), pp. 241-252. https://doi.org/10.1061/(ASCE)LM.1943-5630.0000127

A.A. Enshassi, L.Abuhamra, S.Alkilani, Studying the benefits of building informa-tion modeling (BIM) in architecture, engineering and construction (AEC) industry
in the Gaza strip (2018) Jordan Journal of Civil Engineering, 12 (1), vol. 2, 1, pp. 87-98, http://hdl.handle.net/20.500.12358/26583

X. Liu, X.Wang, G.Wright, J.C.P.Cheng, X.Li, R.Liu, R. A state-of-the-art review on the integration of Building Information Modeling (BIM) and Geographic Infor-mation System (GIS), (2017), ISPRS International Journal of Geo-Information, 6 (2),https://doi.org/10.3390/ijgi6020053

J. Chalhoub, S. K. Ayer, Using Mixed Reality for electrical construction design communication, Automation in Construction, vol. 86, (2018), pp 1-10, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2017.10.028

K. El Ammari, A. Hammad, Remote interactive collaboration in facilities manage-ment using BIM-based mixed reality, Automation in Construction, vol. 107, (2019), 102940, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2019.102940

P.S.Dunston, X.Wang, A hierarchical taxonomy of aec operations for mixed reality applications (2011) Electronic Journal of Information Technology in Construction, vol. 16, pp. 433-444. ISSN: 1874-4753

R. Proctor, and H.Van Zandt Human Factors in Simple and Complex Systems, (1994), Allyn & Bacon

M.A. Muhanna, Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions, Journal of King Saud University - Computer and In-formation Sciences, Volume 27, Issue 3, 2015,

Pages 344-361, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2014.03.023

D.A.Bowman, D.Koller, L.F.Hodges: Travel in Immersive Virtual Environments: An Evaluation of Viewpoint Motion Control Techniques. In: Virtual Reality Annual International Symposium, pp. 45–52. IEEE Press, New York (1997) https://doi.org/10.1109/VRAIS.1997.583043

D.A.Bowman, D.B.Johnson, L.F.Hodges. Testbed Evaluation of Virtual Envi-ronment Interaction Techniques. In VRST '99: Proceedings of the ACM sym-posium on Virtual reality software and technology, pp. 26–33. ACM, 1999. https://doi.org/10.1162/105474601750182333

D.A Bowman and L.F Hodges, Formalizing the Design, Evaluation, and Appli-cation of Interaction Techniques for Immersive Virtual Environments. Journal of Visual Languages and Computing 10, 37–53 (1999)

I.D, Tommelein, and S.Gholami, Root Causes of Clashes in Building Information Models. (2012) IGLC 2012 - 20th Conference of the International Group for Lean Construction.

S.Zhang, K.Sulankivi, M.Kiviniemi, I.Romo, C.M.Eastman, J.Teizer, BIM-based fall hazard identification and prevention in construction safety planning, Safety Science, vol. 72, (2015), pp. 31-45, ISSN 0925-7535, https://doi.org/10.1016/j.ssci.2014.08.001

P.Raimbaud, F.Merienne, F.Danglade, R.Lou, J.Hernandez, P.Figueroa, Smart Adaptation of BIM for Virtual Reality, Depending on Build-ing Project Actors' Needs: The Nursery Case, 667-668, (2018), https://doi.org/10.1109/VR.2018.8446288

P.Raimbaud, R.Lou, F.Merienne, F.Danglade, P.Figueroa, J.Hernandez. BIM-based Mixed Reality Application for Supervision of Construction. 1903-1907, (2019), https://doi.org/10.1109/VR.2019.8797784

P.Raimbaud, M.B.Palacios, J.P Romero Cortes, P.Figueroa, R.Lou, F.Danglade, F.Merienne, J.Hernandez, A Virtual Reality and BIM Approach for Clash Resolu-tion, Accepted, In press

INTERACTIVE URBAN SPATIO-TEMPORAL DATA EXPLORATION TOOL: A WEB APPROACH

MIGUEL FEIJOO, JOSE TIBERIO HERNANDEZ

IMAGINE TEAM

SYSTEMS&COMPUTING ENGINEERING UNIVERSIDAD DE LOS ANDES

ABSTRACT

The analysis of spatio-temporal heterogeneous data is considered as a big challenge due to the emerging of vast quantities of such information that considers the resolution of the questions “what, when and where”. Thus, in urban planning such analysis, through visual analytics, represent a big yearn to experts to get to know perhaps hidden and valuable information, supporting their decision making.

This emerging field has various investigations which are briefly mentioned in order to get an idea of what is currently done and how. In this order of ideas, experts on the field, through this brief approximation, can get a notion of how the data is structured and, through the very best visualization to represent the current data selection of interest, regarding spatio-temporality, get an overview of the dataset as a whole. This article presents a web approach to offer exploration tools of spatiotemporal data.

This proposal is illustrated by a first prototype to show its technical feasibility and, a review of differents proposals is presented and how, based on its analysis, is structured this proposal and its future evaluation.

Index Terms: Human-centered computing; Visualization techniques; Visual analytics; spatio-temporal data;

1. INTRODUCTION

The heterogeneous spatio-temporal urban data is a nowadays challenge for the analysis and decision making through its visual representation.

This conception is due to the quantity of information hidden in datasets with structures that represent events which describes a spatial and temporal phenomenon that exists at a certain time and location, with

the yearn of analysis tending to be predictable and descriptable. The speedy growth of spatiotemporal datasets due to broad acollection of networks and locationaware decides has boosted the demand in spatiotemporal data analytic approaches.

This first approach takes into consideration the analysis related to the exploration and quality of urban data in the capital of Colombia, Bogotá DC. Its revision will be carried out through visualizations that allow the experts to conceive the data without effort, giving them a global idea about how well they are and how they are structured. This brief study will make the reader understand the procedures to reach the leading solution, about the processing and exploration of the datasets provided by the entities of the District, which among them there is heterogeneity, with data both georeferenced and non-georeferenced.

This current study will be guided by two main aims. Firstly, this proposed tool wills to allow the connection to a spatio-temporal data source that permits to make selections of the variables of interest, coupled with the best approximation of visualization that can be changed from one set to another effortlessly. Is yearned to easily allow the exploration of a dataset, considering its heterogeneity, through the selection of variables of interest, which will enable an expert to define new insights from a specific analysis. Moreover, the solution generates a gadget through the saving of the addition of a selection, the best outgoing visualization for it result from spatial choices over the display and possible annotations made by the expert about its findings.

On a second big phase, as soon as is completed this first approach, must be executed some simulation models, arising when data are collected across time as well as space and has at least one spatial and one temporal property. This permits the expert to get to analyze its results accurately, taking into consideration the comparison between historical and simulated or projected data, through different scenarios, in order to get to decision making in an effortlessly way.

Through a literature review we get to know which studies have been done lately regarding about data exploration, types of analysis in the Spatio-Temporal domain and Data Quality, about what it is and how they perform it. On the other hand, in this first approximation, will be described the process of the understanding of the data provided through its abstraction, following the framework proposed by Tamara Munzner[3], which is lately briefly described in the chapter presented below.

However, that this first phase of the study expects an individual exploration of each data source with a quality attribute and there is no pretension of analysis other than data quality and data exploration.

In order to get to this preliminary solution, the use of Navio, as a tool considered as a widget for summarizing, exploring and navigating multivariate datasets, is desired in this present study. What is yearned is to get a solution in which is embedded the implementation made by Guerra and his co-workers[2], as a widget, that permits the selection (zooming and filtering) as a result of the exploration and navigation through this preliminary tool. Is willed to let the expert have a notion of how the data is, through the very best visualization to represent the current data selection of interest, in spatio-temporal terms, as an overview of the dataset as a whole.

In this paper will be discussed how the management of heterogeneous data effortlessly can be easily worked out by District's entities that yearn decision making through rapid well-done data analysis, as a consequence of exploration and data quality. The above, determined by different visualization techniques, which are the result of data types selection and both historical and simulated reports with annotations through both quantitative and qualitative flexible and interactive analysis.

2. RELATED WORK

In this section, will be described what has been worked lately about Data Exploration and Data Quality, principally, in general terms and as also regarding SpatioTemporality, specifically. That is, will be briefly described what is and how these works pose Data Quality, and for the Data Exploration towards Urban Planning, which are the types of analysis usually made.

The very first step must be to reach what Data Quality means correctly. Following the preliminary conceptual framework mentioned in the work of Wang and Strong[4], the quality of the data refers to the accessibility, interpretability of the data, and the relevance and accuracy of it. That is, each category targets a set of dimensions towards believability, objectivity, completeness, traceability, the variety of data sources, value-added, timeliness, ease of operation, flexibility, ease of understanding, representational consistency and concise representation.

Additionally, depending on the attributes or dimensions treated within a data source, fourprincipal categories results: (1) Intrinsic, denoting that data have quality in their own right, (2) Contextual, highlighting the requirement that data quality must consider within the context of any specific task, (3) Representational and (4) Accessibility, emphasizing the importance of the role of systems. That is that high-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumer.

On the other hand, towards the analysis and exploration of big amounts of data that is usually yearned, a visualization widget for summarizing, exploring and navigating multivariate datasets was implemented in order to achieve this current challenge. That is, Guerra introduced and evaluated the tool called “NAVIO”, which displays full summaries, allows sorting and filtering on ranges and values, and keeps a visual trail of the queries, allowing users to navigate and explore the data effortlessly.

As Guerra mentions in his paper [2], Summarization, referring to getting a general idea of the dataset as a whole, Navigation, and Exploration, involving the availability to run specific desired queries, are the three main tasks that Navio addresses.

It is stated how Navio displays a full summary of the dataset right from the start, displaying missing values, patterns, and distributions, giving the user a basic notion of the data completeness in a effortlessly way. Users then can explore the data while focusing on areas of interest by performing dynamic queries through in visualization set of selections, providing the trail of what the users perform. These set of characteristics are usually missing in the type of tools such as Tableau.

Additionally, to also support users without programming training, is presented “SHIPYARD” as a contribution where users can drag and drop their data and get a Navio visualization for understanding and exploring it, with the availability of setting up the variables or attributes involved by how the user desires.

Notwithstanding these previous preliminary advantages, as running on a web browser as a characteristic of flexibility, both Shipyard and Navio gets to have low scalability as a result, in comparison with such tools as Tableau, regarding their limitation on the support of significant bigger datasets, usually over 400Mb.

Among a large group of commercial tools that support visual analysis, such as Infozoom or Tableau, the latter is the best known and common to achieve these visual tasks easily. Due to its nature, is the user who decides, among the dimensions or values, to visualize dragging and dropping each element on a canvas. As stated in Guerra’s work, this tool is great for the navigation, but it doesn’t lend easily to summarization

or exploration. Being a common commercial tool, the tool is intuitive when you have knowledge of what you are looking for, emphasizing its advantages over navigation tasks, supporting big amounts of data and processing them with a high efficiency.

Current works with variety of temporary query tools such as “TEMPEST”[9], allows the user to select arbitrary combinations of months, days and times of day and see what happened at this time. Furthermore, another type of interactive time filter in the “STRAD”[10] project where the user selects the period of time that he wishes to see by means of a slider where the consultation period is specified, placing the start and end date and time, allowing to visualize on the map the trajectories of ships that moved during that time.

“TrapezoidBox”[8] is a reference tool for the spatial query, implemented on Google Maps, in which the spatial proximity queries are made by means of a trapezoid, where users can drag the four vertices to change the query condition and the result is seen in the map by the circular regions which represent the ranges satisfactory distance from the place of interest.

However, the comparison of spatial data in different instants of time aims to show the differences or proportions between the values for each moment and the values for the previous moment or at any time selected by the user.

For its part, Andrienko (2003)[7] proposes a topology based on the classification and evaluation of how such spatio-temporal data can help to resolve questions through exploration and the characteristics of the data they are applicable to. The proposal directly relates tasks to components of data, in terms of space (where), time (when) and objects (what). That is, questions can be easily answered keeping only what users need to satisfy the query constraints, on what is called lookup and filtering.

On the other hand, Chen et al. (2018)[6] proposes a Visual Analyzer for Urban Data called “VAUD” which supports the visualization, querying, and exploration, allowing the Multi-source analysis by leveraging spatial-temporal andsocial interconnectedness features, and selecting, filtering and aggregating across multiple data sources, which permits the extraction of information that would be hidden to a single data subset. Coincides with Andrienko when defining the resolution of a question on a query as the combination of four main component constraints, denoting the (1) which: Identification Attributes, (2) the when:

Temporal attributes, (3) the where: Spatial attributes and (4) the what: Descriptive attributes. Moreover, is claimed the importance of visual querying, in order to engage more non-expert users.

In this order of ideas, even the on-thefly queries and association of attributes are supported, VAUD requires users that must have a notion of networks, regarding the manipulation of nodes and the specifying of the conditions, to get to explore by zooming, panning and detailing the desired results.

Furthermore, Doraiswamy et al. (2018) [5] states that visual analytics systems such as their tool proposed as “URBANE”, aim to empower domain experts to explore multiple data sets, at different time and space resolutions. In this proposal navigation and operations on map view such as panning, zooming, and rotating the view are accomplished through mouse interactions for analyzing multiple sources.

Speaking in terms of storage, querying and analysis both “VAUD” and “URBANE” softwares achieves a high efficiency. Different alternatives to this are thought considering each software’s context and nature.

To finalize, will be briefly shown in the table below (Table 1) a brief summary of the couple of softwares mentioned above, in terms of (1) Storage, (2) Flexibility, (3) Accessibility,

(4) Querying & Analysis, and (5) Main SpatioTemporal Interactions.

	VAUD ***Chen et al. (2018)***	**URBANE** ***Doraiswamy et al. (2018)***
Storage	To enable cross-domain analysis by leveraging the spatio-temporal interconnectedness, they build a sequence of STCs for spatio-temporal objects. The average memory consumption of an STC is 5Gb. Therefore, the total consumption for 22 STCs is about 110Gb. They store all STCs individually in the hard disk and construct a spatiotemporal index structure to accelerate the online query.	**Raster Join** Being a 3D Map proposal, a rasterization-based approach is thought to leverage current generation graphics hardware (GPUs), storing the different urban data sets in a 3D grid index of fixed size, where the dimensions correspond to the location (2 coordinates) and time.
Flexibility	**Multi-source Analysis / Interactivity** * Manipulating Nodes: The analyst can create a node by moving a node onto the query view. * Specifying Conditions: The analyst sets a query condition by first adding a node in the query view and then specifying the detailed conditions.	**Multi-source Analysis /Interactivity** Urbane generates queries for two different operations—visualizing on the map, and visualizing on the PCC [Exploratory View]. For both these cases, they execute Raster Join using a pre-configured 20 meter bound. However, users can change this bound if they require higher accuracy.
Accessibility	**NO** WEB	**NO** WEB
Querying & Analysis	**Multi-source Analysis** * Exploring Results: The analyst is able to select one or more objects from the result node and place these in the scene view. The analyst can pan and zoom to explore details in the scene view. Furthermore, an analyst can explore detailed information by clicking an object.	**Multi-source Analysis** The main goal of the data exploration view is to support the analyses of urban data at two different resolution levels—region and building. **Exploratory view** This visual representation is effective for analyzing multivariate data, and can provide insights into the relationships between different indicators. Users can also filter regions by brushing the desired range of values on individual axes of the PCC. This updates the map by highlighting all regions that satisfy the filter constraints

	VAUD ***Chen et al. (2018)***	**URBANE** ***Doraiswamy et al. (2018)***

Main Spatio-Temporal Interactions	**Space-time-cube based (STC)** Constructed an STC for each time slice and uniformly subdivide the STC into a 3D grid for a given resolution, where the resolution is determined based on the analysis tasks. As such, a cell of the STC refers to a geographical location and a time interval in the time slice associated with the STC. Finally, we sequentially relate each record of each object into an STC cell by leveraging the timestamp and location information. The spatio-temporal data and associated STCs support fast querying of spatiotemporal information and facilitate indirect connections of objects by means of the spatial-temporal interconnectedness.	**Map View** This view is composed of a map rendering component. The various menus and panels are overlaid on the map. Navigation and operations on map view such as panning, zooming, and rotating the view are accomplished through mouse interactions. The main menu (right side of map view in Fig.1) allows users to control all the functionalities of the system. This includes loading or deleting urban data sets as well as polygonal regions that define the different resolutions. Users can then choose the data set to be visualized along with the visualization resolution. Multiple spatial aggregation queries can be generated based on the user interactions; thus providing efficient support for these queries is crucial.

Table 1 Comparative Analyss Between Spatio-Temporal Softwares

3. METHODOLOGY

An iterative development strategy will be adopted based on experiences with data and users increasingly involved in the process, such as the pre-processing and procedures followed approximately by the authors in the previous related work.

That is to say, in the first place the understanding of the spatiotemporal content structures is proposed, in order to proceed with the adequate analysis for them, by solving the set of possible questions among the attributes on the WHAT, the WHEN and the WHERE, to carry out a successful analysis and understanding of this data. Such tools proposed above, coincide with the general need for visual interaction on a work canvas (i.e. maps or networks), in order to achieve in the user understanding of data without the need of technical knowledge in computing or database management, particularly.

In our terms, as the aim of this brief study is to get to analyze the information towards Urban Planning in the city of Bogotá D.C. in Colombia, was provided a set of datasets which, preliminarily, must get to be analyzed. For achieving this first requirement about the understanding of the data presented, will be described as detailed as possible, the abstraction of it following the Framework of Tamara Munzner [3].

This framework let us focus on tasks and effectiveness, serving as a constraint on design, allowing us to get to understand a unique common language, avoiding domainspecific terms. This data abstraction follows three main significant parts: WHAT, WHY and HOW, following this order preferably. The reason why the order of analysis matters is due to the responding of these leading questions that perhaps should be

described as: (1) WHAT (data abstraction), (2) WHY (task abstraction) and (3) HOW (information about the idiom and its interactions).

To get to analyze all these three requirements, willing the best visualization and interaction to answer all the tasks proposed and planned, must be followed this next procedure. In the data abstraction, should be discriminated the Datasets and its Attributes. Following this, all data types and dataset types, as also all its attributes types should get identified. Furthermore, is also essential to determine the availability of every dataset provided, to get to the visual solution.

As soon as the data is all understood through its abstraction, is critical to get to know all “actions” versus “targets” which grant the recognition of the tasks that the collection of visualizations must follow, through three primary levels: Consume, Search and Query. Finally, as the data and tasks abstraction are identified, is vital to get to know the complete encoding of each visualization, determining the best masks and channels following the principles of expressiveness and effectiveness, through their rankings.

Given the above, there is preliminary need to address a first mono-source phase with a basic set of quality features, in order to achieve the heterogeneity between different single-sources and its fulfillment on the data characterization towards spatiotemporality. Due to that, is first yearned to get a solution in which is embedded the implementation made by Guerra: NAVIO, as a widget, that permits the selection (zooming and filtering) as a result of the exploration and navigation through this preliminary tool.

Is willed to let the expert have a notion of how the data is, through the very best visualization to represent the current data selection of interest, as an overview of the dataset as a whole. That is, a solution where through a single or group of selections that the expert/user performs over the embedded widget of Navio, let display the very best visualization over the attributes selected, that allows the user to get a general or even specific notion of the quality of the data provided to the tool (such as outliers). The implementation will be made by using the library D3.js, Vega-lite, DataTables of jQuery and JavaScript.

4. CURRENT WORK

To achieve that need of effortless on the regular analysis of big datasets, in the very first one will be carried out a single dataset analysis through an Interactive Urban Spatio Temporal Data Quality Exploration tool. Regarding the visual yearn, will be briefly applied the principles on visualization following the framework of Tamara Munzner, presenting the very best approximation of a visualization between two variables, as allowing the user the flexibility through its selections, heterogeneity among the different sources to analyze and visual analysis decisions.

In the second approach, as most visual analytics tools tend to focus only on a single data source, making it difficult to discover and link overlapping details of an event from multiple data sources, will be carried out a multi-source analysis and, the application and analysis of simulation models for decision making.

As told before, this first approach is toward the very first need, which is to get a solution based on an interactive urban Spatio-temporal data quality exploration solution, displayed on the web.

That is, we propose the implementation of a web-based tool towards the exploration and analysis of the data quality of one set of spatiotemporal urban data, preliminarily for this first phase, which permits the user to get a general notion of how the data is structured and registered. Thus, having a web access to the proposed tools is a self-imposed requirement, turning to the general statement of “everytime, everywhere,

but not everyone" that this tool supports. That perception of the expert resulted through the tool, is the outcome of an automated preprocessing which organizes the current data presented in a dataset. Additionally, of how the interactive querying in the widget of NAVIO and of how the user selects its desires supported by the very best visualization following the framework proposed by Tamara, linking the best combination between categorical, scalar and Spatio-temporal attributes.

The application that is presented right away turns over the need of specification towards the three main aims of spatiotemporal analysis, regarding the WHAT, WHEN and WHERE, interactively. In this first phase of the preliminary study, is pretended to show the user a flexible interactive tool on where visual queries can be done all over three main displays that helps answering these questions at particular selections desired all over the navigation and exploration supported.

This interactive tool permits the interaction all over time and space towards answering the main questions, as mentioned before in the literature review, particularly in the topology claimed by the work of Andrienko.

In this order of ideas, the interactive urban spatio-temporal data exploration tool proposed achieves a global understanding of the three main big questions that should be asked when talking about spatio-temporality.

That is, in the Figure 1 is presented the visual definition of the WHAT taking as main input the widget "NAVO". Moreover, is presented the WHEN, starting from the idea of comparing attributes in quantitative terms over the time (yearly, monthly and even daily), as can be appreciated in the Figure 2. Furthermore, is presented the WHERE towards the understanding of the space and its influence over the analysis, as shown in the Figure 3.

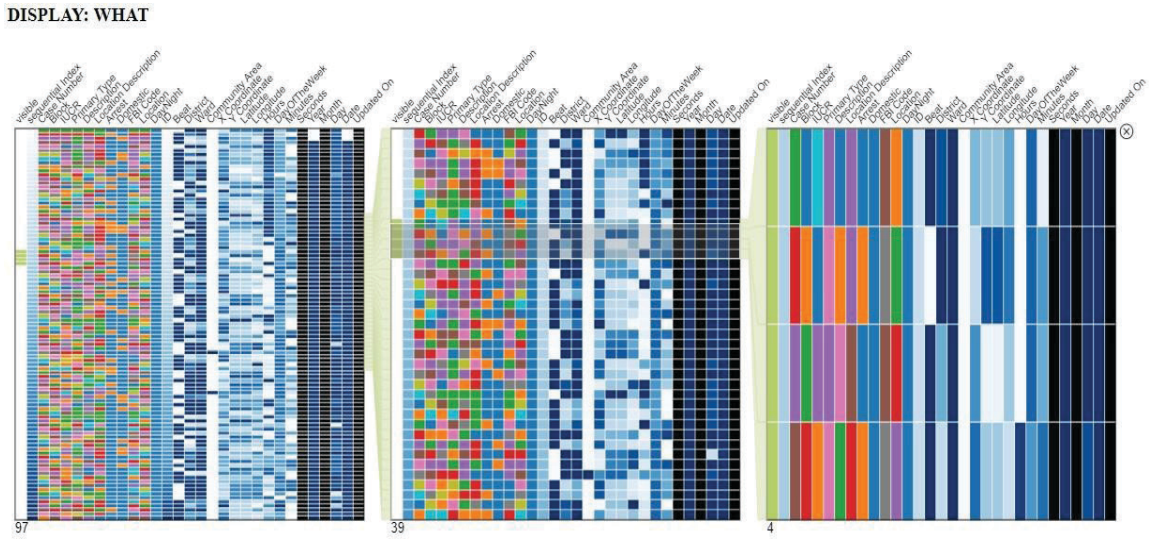


Figure 1 Visual Definition – WHAT / NAVIO

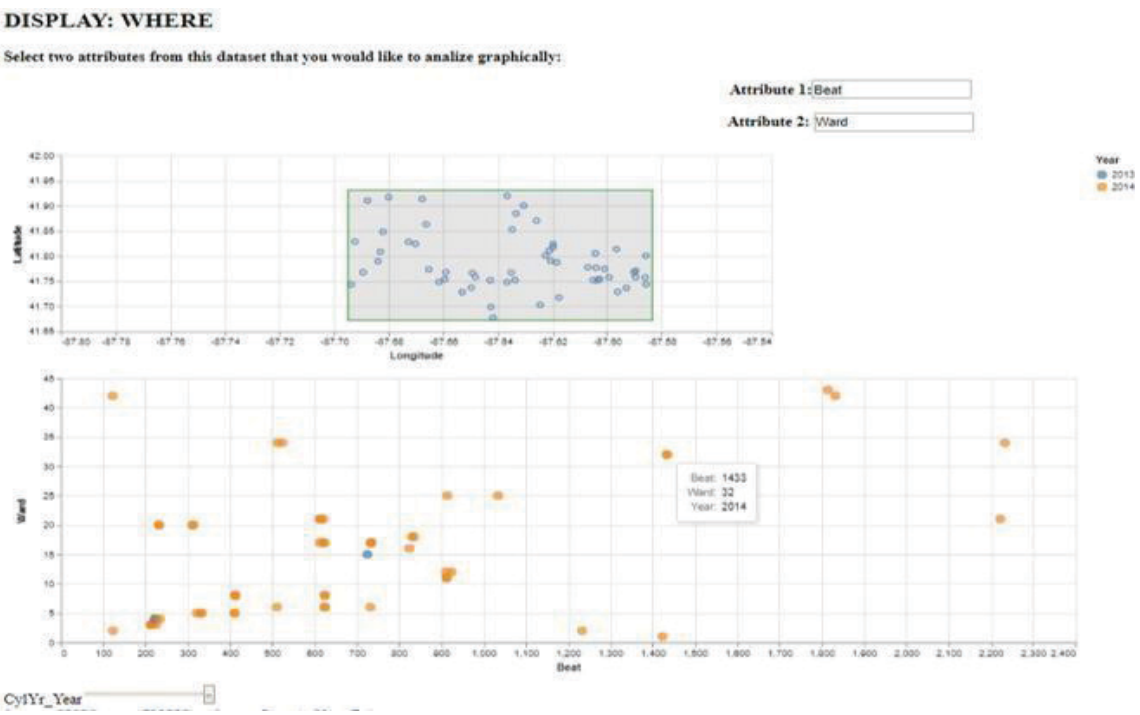


Figure 2 Visual Definition – WHERE (SPATIAL ANALYSIS)

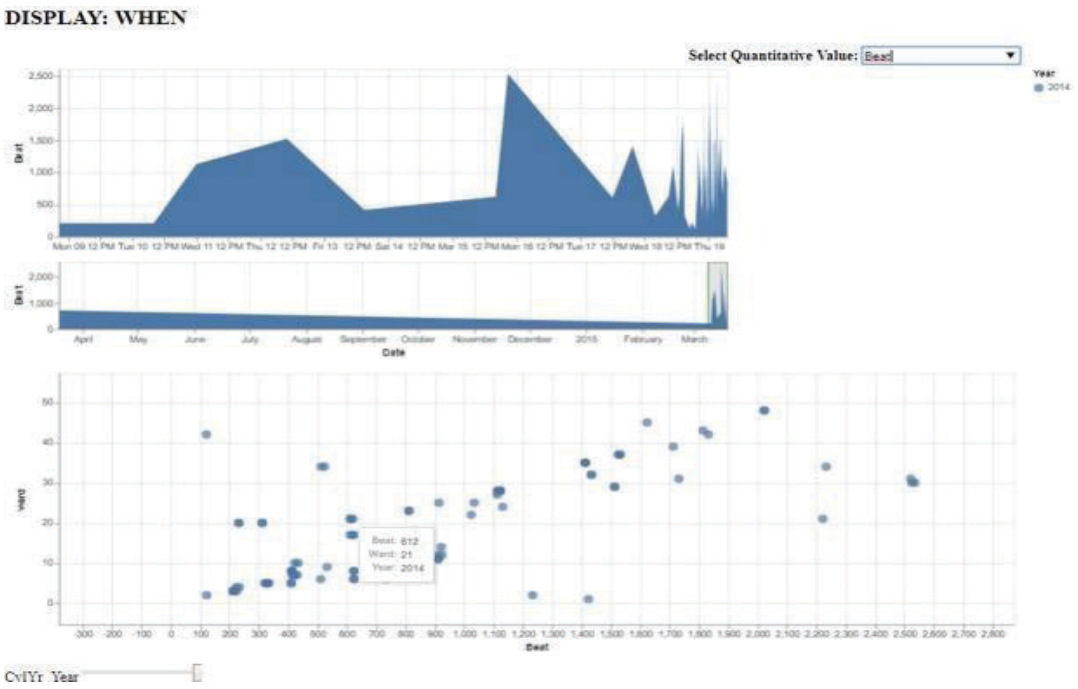


Figure 3 Visual Definition – WHEN (TEMPORAL ANALYSIS)
a search field. This same idea is supported by every

Notwithstanding this preliminar solution is taking into consideration a monosource analysis, is important to emphasize that it is a first phase in which it is intended to show the interaction between space and time for a set of data, to achieve a much more disaggregated analysis, based on the idea of flexibility in the selections or conditions of interest, that this tool is pretended to support.

Filtering on brushing, selections, Scrolling, Panning and Zooming are permitted in the proposal to get an analysis much more detailed. Summarizing, Comparing, Analyzing, Navigating, are the main tasks fulfilled in this first preliminary resulting tool, which can be accessed entering to the following link: https://goo.gl/WVSZpf

On the other hand, towards the need of registers of a specific query all over the data displayed, the tool provides a paginated table with characteristics that show the visible data over the selection on widget of Navio, that also permits the user to get an export of that data or even find more specifications on visualization displayed on the web-based solution, allowing the user to save as wanted selections, brushes and filters, respecting in this way, the Shneiderman's visualization mantra of "Overview first, zoom and filter, then details-on-demand", which dictates that a visualization systems should provide a good overview of the data, support zoom and filtering options, and offer details when the user requests them. In the Figure 4, presented below, is shown a sample of what was mentioned above.

Following the previous mentions, is must be emphasized that this preliminary result of software implemented, has characteristics of flexibility, interactive querying on the widget and visualizations. That is, the fulfillment of what was pretended roughly was accomplished with results not only in terms of spatio-temporality review all over the related work, but with a functional prototype for the tasks previously destined for it.

Technical tests were made to guarantee the operation and flexibility of the system for the analysis over different data sources, with variation all over its structure and attributes. In terms of scalability on the charging of big amounts of data in a dataset, there is a limitation due to its reading and "calculations" or interaction on the web browser, that stucks on a maximum of 400MB, even lower. However, regarding the supporting of heterogeneity desired over, tests were made over two datasets achieving the correct reading and interactions of their data within every visualization. Results were how it was willed in a general manner, considering the aims of this first stage. That is, considering every possible typology or structures of every dataset's attributes, a preprocessing of every possibility (regarding time and space, particularly) was made, so that no matter which dataset is charged to get its visual analysis, the fields concerning spatial and temporal attributes will always be the same.

Copy CSV Excel PDF Print Search:

ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domest
10000092	HY189866	Tue Mar 18 2014 22:44:00 GMT-0500 (hora estándar de Colombia)	047XX W OHIO ST	041A	BATTERY	AGGRAVATED: HANDGUN	STREET	false	false
10000094	HY190059	Tue Mar 18 2014 23:00:00 GMT-0500 (hora estándar de Colombia)	066XX S MARSHFIELD AVE	4625	OTHER OFFENSE	PAROLE VIOLATION	STREET	true	false
10000095	HY190052	Tue Mar 18 2014 22:45:00 GMT-0500 (hora estándar de Colombia)	044XX S LAKE PARK AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	false	true
10000096	HY190054	Wed Mar 18 2015 22:30:00 GMT-0500 (hora estándar de Colombia)	051XX S MICHIGAN AVE	460	BATTERY	SIMPLE	APARTMENT	false	false
10000097	HY189976	Wed Mar 18 2015 21:00:00 GMT-0500 (hora estándar de Colombia)	047XX W ADAMS ST	031A	ROBBERY	ARMED: HANDGUN	SIDEWALK	false	false
10000098	HY190032	Wed Mar 18 2015 22:00:00 GMT-0500 (hora estándar de Colombia)	049XX S DREXEL BLVD	460	BATTERY	SIMPLE	APARTMENT	false	false
10000099	HY190047	Wed Mar 18 2015 23:00:00 GMT-0500 (hora estándar de Colombia)	070XX S MORGAN ST	486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	false	true
10000100	HY189988	Wed Mar 18 2015 21:35:00 GMT-0500 (hora estándar de Colombia)	042XX S PRAIRIE AVE	486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	false	true
10000101	HY190020	Wed Mar 18 2015 22:09:00 GMT-0500 (hora estándar de Colombia)	036XX S WOLCOTT AVE	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS	STREET	true	false
10000104	HY189964	Wed Mar 18 2015 21:25:00 GMT-0500 (hora estándar de Colombia)	097XX S PRAIRIE AVE	460	BATTERY	SIMPLE	RESIDENCE PORCH/HALLWAY	false	false

Showing 1 to 10 of 97 entries Previous 1 2 3 4 5 … 10 Next

Figure 4 Paginated DataTable / Queryable& Exportable Characteristics

Regarding the visualization obtained over every single dataset, in this preliminary phase, the user will be having a notion of how the data is structured (i.e. Non-Values on a specific attribute) that could affect

the analysis. On the other hand, the user will have the tools to obtain preliminary insights towards the data at specific time or space, as also the possible outliers or highlights that could be made in this interactive proposal.

CONCLUSIONS & FUTURE WORK

As told before through the whole paper, the proposed tool is flexible that allows effortless analysis on the Web. On the other hand, other types of data, the temporary spaces represent a challenge of superior analysis, for decision making, particularly in urban terms, from the recurring questions space (where), time (when) and objects (what).
What is willed in a second phase of the present study it is proposed that as soon as this first approach is completed, some simulation models should be executed that allow the expert to analyze their results with precision, considering the comparison between historical and simulated data or projected, through different scenarios, to decide without effort. Additionally, it is proposed to add multi-source reading and analysis in this tool proposed.

REFERENCES

Batini, C., Barone, D., Cabitza, F., & Grega, S. (Febrero de 2011). A DATA QUALITY METHODOLOGY FOR HETEROGENEOUS DATA. International Journal of Database Management Systems, 3(1).
Guerra Gómez, J. A., Ortiz Román, J. C., & Murillo Castillo, J. G. (2018). Navio: a visualization widget for summarizing, exploring and navigating large multivariate datasets [En revisión] Munzner, T., & Maguire, E. (2015). Visualization Analysis & Design. FL: Taylor & Francis Group. Obtenido de https://bit.ly/2FHJJxE
Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy:What Data Quality Means to Data Consumers. (T. &. Ltd., Ed.) Journal of Management Information Systems, 12(4), 5-33.
Doraiswamy, H., Tzirita Zacharatou, E., Miranda, F., Lage, M., Ailamaki, A., Silva, C. T., & Freire, J. (2018, May). Interactive Visual Exploration of SpatioTemporal Urban Data Sets using Urbane. In Proceedings of the 2018 International Conference on Management of Data (pp. 1693-1696). ACM.
Chen, W., Huang, Z., Wu, F., Zhu, M., Guan, H., & Maciejewski, R. (2018). Vaud: A visual analysis approach for exploring spatio-temporal urban data. IEEE Transactions on Visualization & Computer Graphics, (9), 2636-2648.
Andrienko, N., Andrienko, G., & Gatalsky, P. (2003). Exploratory spatiotemporal visualization: An analytical review. Journal of Visual Languages & Computing, 14(6), 503-541.
Burigat, S., & Chittaro, L. (2005). Visualizing the results of interactive queries for geographic data on mobile devices. Proceedings of the 2005 International Workshop on Geographic Information Systems - GIS 05.
Andrienko, N., & Andrienko, G. (2004). Interactive visual tools to explore spatiotemporal variation. Proceedings of the Working Conference on Advanced Visual Interfaces - AVI 04.
Fernandez-Prieto, D., NaranjoValero, C., Hernandez,
J. T., & Hagen, H. (2017). STRAD Wheel: Web-Based Library for Visualizing Temporal Data. IEEE Computer Graphics and Applications, 37(2), 99-105.

RICKSHAW MANAGEMENT FLEET SYSTEMS BASED ON IOT AND ITS APPROACHES FOR LAST MILE TRIPS

LUIS ANDRES MARENTES1, LUIS FELIPE HERRERA-QUINTERO1,2, DIEGO BERMUDEZ1,2, AND JULIAN VEGA1

1 MINISTRY OF TRANSPORTATION, COLOMBIA HTTPS://WWW.MINTRANSPORTE.GOV.CO/

2 PILOTO UNIVERSITY OF COLOMBIA, BOGOT´A HTTP://WWW.UNIPILOTO.EDU.CO

ABSTRACT.

Since the beginning of the 18th century the bicycle has been used daily and it has become an alternative mean of transportation for cars, thanks to its reduced cost, ease of use, size, positive influence on the people's health, low environmental impact, and time-efficient travels. Moreover, the bicycle has been adapted to develop public transport services of cargo or passengers using among others, assisted pedaling, which is known as rickshaw vehicles.

This context has led to the development of road infrastructure in countries such as Denmark, Holland, Colombia, India, among others, to make public places better adapted for cycling, offering greater security, through exclusive lanes and, through these programs, finally, reduce the use of the car and increase the use of this sustainable and environmentally friendly transport alternative. As from the above described, use cases for rickshaw as a public transport mean have created new challenges for countries and their cities at the light of stakeholders (drivers, users, authorities, etc.); one of them is fleet management for rickshaws in cities. For fleet systems, a diversity of technological solutions might support the context of rickshaws.

One of them is using the concept of Internet of Things (IoT) that essentially addresses the deployment of small devices of low consumption and low cost. In this way, this paper proposes a novelty solution based on IoT communications that allows government entities to monitor the minimum safety conditions for the service, including the route and speed enforcement. The proposed solution takes advantage of very low cost IoT devices reaching minimum investment and operation for drivers and provides information to travelers.

Keywords: Smart Mobility · IoT · Smart Cities · ITS

INTRODUCTION

Informal transport services, defined by [3] as "paratransit services provided without an official sanction", have been developed around the world to supply uncovered demand. Services using pedicabs, e-bike taxis, or rickshaws are examples of informal transportation services. These non-motored vehicles have been adapted to perform small trips, less than a mile, and offer a service, which is by nature, route flexible, quick, and low-cost.

There are benefits and social costs resulting from pedicab operation. To serve the objetives of

municipalities, the service provided by pedicabs should be detail controlled, so the service become a solution instead of a problem. Regarding benefits, it has been proved that their operation can serve as a way to fulfill market inefficiencies in places unserved, as last mile corridors. In fact, as long as cities have been developing public mass transportation infrastructures, there have appeared small sectors where people find useful to use pedicabs as a way to reach the public system, [18] and serves as a feeder for mass transportation system. In fact, pedicabs operators are normally quicker to respond and adapt to changes in demand than formal transportation services and can operate with relative lower costs.

In [3], authors explain as a first problem the common practices of erratic scheduling and supply rationalization. In most of the cases Pedicabs routes run as a monopoly of a cooperative of drivers. Therefore, the cooperative maximizing profits might make headways become very long for users, so the user satisfaction is low. A second problem is the lack of accountability and evasion of other legal regulations, these two issues have, as a consequence, that there is no way to complaint about the service. Moreover, as users are dependent on the route, then the cooperative or drivers have the relationship power. Finally, a third problem is related to the lack of control on driver's behavior on routes. Cooperatives or drivers running these services not only evade taxes, but also other regulations; such as, minimum salaries, working hour restrictions, age limit for drivers.

Despite of the benefits and social costs exhibited by pedicabs as one of the least intrusive forms of movements [2], and the well known fact of local governments creating special infrastructure for incentivizing bycicle trips and by this way pedicabs trips too; for instance, some Colombian, Dutch and Danish cities; there is a growing concern regarding their safety operation. To the best of our knowledge safety studies on pedicabs has not been published, but there are on bike and e-bike riding that may serve as a reference.

Most of the safety research is evolving around two topics (1) conflict and aberrant behavior mainly at intersections and crash and injury data from hospitals and crash records. Regarding the first topic, e-bike riders seem to violate more rules that any other actor on the road; in fact, authors of [19] found that in China crossing in red-light count for 56% depending on demographic factors, A similar study in Melboure, Australia [7] found that 5% violate the traffic light. Other factors like using phones while riding, helmet use, riding in the wrong way or outside the bicycle lane have been studied, in [1] authors observe that 25% of the ridders violate some of these rules. Moreover, there are some concerns about the possibility of mixing assisted pedaling and normal bycicles in the same line. Many studies reported, [4], [10], [20], that e-bikes travel speed is 40%-50% faster and this fact might produce more safety problems on the road.

Therefore, to take advantage of pedicabs as feeders of mass transportation systems, control their operation regarding the quality of the service being delivered and make sure that their drivers follow safety rules, there is a need for a surveillance solution. This paper presents such a solution based on Internet Of Things (IoT) concepts. In particular, the following are the contributions of this work being developed:

1. We introduce the system requirements to control service and its safety delivery conditions.
2. We present a networking architecture that uses Lora as the communications technology, so the solution is low cost

The rest of the paper is organized as follows, in section 2 we present the literature review. After that, in section 3 we introduce a set of requirements for the solution. Next, we study our proposal in terms of the networking architecture. Finally, in section 5 we conclude the paper including the following steps to carry out.

2. LITERATURE REVIEW

This section is divided into two parts considering the architecture of previous bicycle tracking systems. The first subsection presents the functionalities of those systems. In the best of our knowledge, there are not any proposals published for pedicab tracking. As a consequence, We search for proposals on architecture and systems for bike tracking in general, and, in a great extent, e-bike tracking as pedicabs are bikes with pedaling assistance. In the second subsection, we review communication mechanisms used. This architecture element is critical for the proposal as implementation cost depends on large extend on the communication technology.

2.1 TRACK SYSTEM FUNCTIONALITIES

There are many proposals for e-bike tracking with different scope, see table 1. All of them follow the bike location. E-bike are capable of assistance during peak effort periods though small engines, a second group of proposals monitor how that assistance is being used. A third group of proposals have been considering battery and the health conditions of riders. In what follows, we review all these proposals resuming their proposed functionality. As the reader can see on the blue table section, none of them have been created to track safety driving conditions as well as service quality, which are our focus.

Tracking systems [5,6,8,9,11,13,16,17] by means of a component sending each bike's location (longitude, latitude, altitude and time) to a central server follow the location. The system normally uses a GPS equipment which sends the information within a pre-specified frequency. This information is used to track in detail the trip and plot it on a map for further analyses and for route sharing.

Functionality	Dill & Bike Active	Static Ubi-Gliebe		Copenhagen wheel	Peafgen & SEM Michahellens	Bike-net Project	Campus Mobility	
Open source Software	No	Yes	Yes	Yes	No	No	Yes	Yes
Hardware Requires	No	No	No	No	No	No	No	Yes
User Interaction	Yes	Yes	Yes	No	No	No	No	No
Wheather Proof	Yes	No	No	Yes	Yes	Yes	No	Yes
Track bike location	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
(GPS) Real-time sensing Track motor assistance Battery	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Power	No	No	No	No	No	No	No	Yes
	No	No	No	Yes	No	No	No	Yes
Traffic light Compliance	No	No	No	No	No	No	No	No
Speed control	No	No	No	No	No	No	No	No
Use Bike Infras	No	No	No	No	No	No	No	No

Table 1. Comparison among different rickshaw implemented systems.

One of the advantages of detail tracking is that servers may take data location records as input for calculating among others the cyclist effort analysis on different routers, bike speed and topography related information found on municipal maps. Normally, users of the monitoring system receive feedback from their route habits, new popular routes suggestions, which create a digital place to share and find the safest, most enjoyable and more efficient route from the city.

A second aspect is to control the motor assistance during pedalling. The reason to build these sets of functions is that cyclists are interested in receiving the assistance in the hardest parts of their journey, probably those with the highest elevation percentage, automatically designated by a sensor, as is the case of [17] or through a direct order of the cyclist as explained by [8]. Sharing the information when the assistance is active not only helps to create assistance profiles on route, but also to understand the user characteristics in order to make possible to optimize them.

A third aspect monitored by those systems is battery consumption. [8] they use the information to detect fleet problems, inform the user of a required recharge of batteries and alert users of this event. In a more service oriented approach, the authors in [14] propose to use the battery information for self-managing the recharge cycle on bicycle docks.

Environmental conditions as well as the health condition of riders are other focus of bike tracking systems. Bike riders are prone to diseases related to air pollution and to disorders caused by noise at streets. So one important area to monitor is that environmental conditions during riding are within acceptable ranges, which makes riding healthy.For instance [17], employs a set of sensors to measure nitrogen oxides, carbon monoxide, temperature, humidity, and noise.

However, measuring the environment is not sufficient, there is the need to control the health condition of the cyclist. Thus, proposals have been published to design applications with the scope of tracking health. In [9] authors explain a system capable of monitoring the physical condition (heartbeat, temperature, calories consumed) of the cyclist and, in case of accident, calls nearby people and report the location for assistance.

COMMUNICATION TECHNOLOGIES AND DEVICES

In terms of the communications technologies, there have been a few different proposals to backhaul the required traffic. The most popular option uses cellular technologies to send the information to the server, [5,6,8,9,11,13,16,17]. The communication is done using 2G, 3G or 4G as the connection technology and assume the use of the cellular phone of the rider or installed on-board. In general, proposals are based on the Android API, which makes possible to extend phone's sensors (accelerometer, GPS) in order to gather additional information from the deployment environment.

Due to concerns around battery consumption by cell phones, there are other authors that recognize the use of cellular networks as a problem and created new options. For example, Authors in [15] propose developing a wireless network to backhaul the traffic. Besides lower battery consumption, there is a cost reduction benefit when using another communications technology.

Indeed, [12] authors show that LPWAN technologies are better suited to track tasks. Lora and Sigfox and NB-IoT, which are the most salient technologies in this group, are best suited to deliver packets on long distances, with low data rates, and cost effectiveness. This paper, as the best of our knowledge takes the lead in introduce Lora to e-bike tracking. This decision is based on the low cost of the equipment required for the target scenario and the low cost of gateways which make the total cost of ownership at the lowest of all three technologies.

SYSTEMS REQUIREMENTS

In the scope of the ITS solutions, we took advance of V methodology proposed by Federal Highway Administration (FHWA). For that reason, we have proposed a concept of operation focused on ITS Service for the users as of the service provide by rickshaws. As regard of this, we consider the follow minimum necessities and expectation that was collected in the scope of stakeholders as follows:

- Speed and route control
- Designated Infrastructure use control
- Bicycle driver behavior control
- Processing requirements on bicycles
- Information delivery to drivers

VEHICLE TERMINAL UNIT - (VTU)

This terminal is made up of an electronic device capable of transmit and receive information and, in the scope of intelligent transportation systems, is called on board unit. In addition, and in accordance with our proposal, this device is belonging to the internet of things solutions because its functionalities are described as follows:

The device must be capable of delivering information about the location of the rickshaws. Speed and sudden stops should also be calculated at the device in order to send only unruly events.

- The device must coordinate its operation with the gateway in order to inform users regarding the location of rickshaws drivers and its availability, in face of the last mile service transportation.
- The device must be provided with image recognition capabilities to detect aberrant behaviors; such as, traffic light violation, riding in the wrong way or outside the bicycle lane.
- The divice must be capable of receiving information to coordinate and control the service provision. A low cost display must be part of the device to show the information.

–

Gateway communication function - (V2I) This part of the solution is very important because all the IoT devices must send information to a gateway. Afterwards, the gateway resends data to a target network. Essentially, gateways establish several links among several networks, it means that, they provide communications to a remote network or an owner system. Gateways work like entry and exit points in a network allowing to gather and share data. We mentioned V2I (vehicle to infrastructure) because this device fulfills several communications functionalities that are used by the general architecture in order to provide the last mile ITS service.

Event collection function (location, velocity, sudden stops, etc) In accordance with the above-mentioned necessities, our solution will provide the information based on events. The central system collecting the information from VTU must process the events generated in real time and be capable of trigger related actions.

Data sending and receiving function As regard to this function it is important to highlight that IoT solutions depend on the way that IoT devices send the information to the gateway and this aspect can be done using several wireless solutions (Bluetooth, Wifi, LPWA – Low Power Wide Area) and different frequencies (433MHz, 868 MHz, 915MHz, 2.4GHz). However, like we described before, the ITS service is focused on last mile scope, therefore many aspects must be taken into account by the data sending function; one of them is the energy consumption of the IoT devices because this factor is crucial either by users or the central systems.

3.2 MANAGEMENT AND OPERATION CENTER

The management and operation center is a piece of the whole system where the information is processed and interpreted by the authorities in order to make decisions. In addition, this part of the systems must send information about several different variables that generate value for the users. At the same time, this center must generate the tools that allow the operations of the rickshaws, for that reason tracking solutions based on map applications should be considered. In light of this, it is required to deliver a center management, a route management, and a data management functions must be considered to fulfill the controlled scenario.

3.3 SERVICE PRESENTATION FUNCTION - (I2P)

Although in this system all the elements are a key point of the solution, the service presentation function is probably the most important part of the equation, that is because the presentation service must provide all the information about the rickshaw and rickshaw-driver through a mobile device. We propose using REST-based approaches and security layers to perform the communication with the server.

4. NETWORKING ARCHITECTURE

The fundamental aim of our proposal is focused on how we can integrate an IoT approach based on LoRA devices, especially because this is a recent approach that can be used like a last mile solution where the rickshaw scenario is deployed. Additionally, we took advantage of this IoT approaches that is used in an indoor or outdoor configuration, however, our proposal will work in outdoor scope. We organize our proposal creating some layers where information or data is gathered and exchanged through them (see Figure 1). These layers are explained below.

4.1 MONITORING LEVEL

This level includes LoRA technologies that are associated with monitoring or watching several physical phenomenon. For this reason, it is important to note that we can use various kinds of sensors. Our target rickshaw includes a 12V battery, so we are taking advantage of this voltage source to power our VTU. In the initial setup, at least it is required a MMA8451 accelerometer, a night vision fisheye camera 5mp IR-CUT automatically switching between day-vision and night-vision, and a GPS L80-M39/L80GR01A07S with nmea format for location data and we are using a 1575.42 MHz GPS antenna in each VTU.

Data collected can be sent in several frequencies (433 MHz, 868 MHz, and 915MHz). Moreover, it is important to monitor and display information received from the gateway, which takes part in the integration level. So LoRa devices should include a small display (0.91 Inch) that works though I2C bus. From the requirements, the proposal must be a Class C Lora device, which has the lowest latency and is bi-directional.

The architecture considers the specification defining the device-to-infrastructure (LoRa) physical layer parameters & (LoRaWAN) protocol. So provides seamless interoperability between manufacturers, as demonstrated using a certification scope.

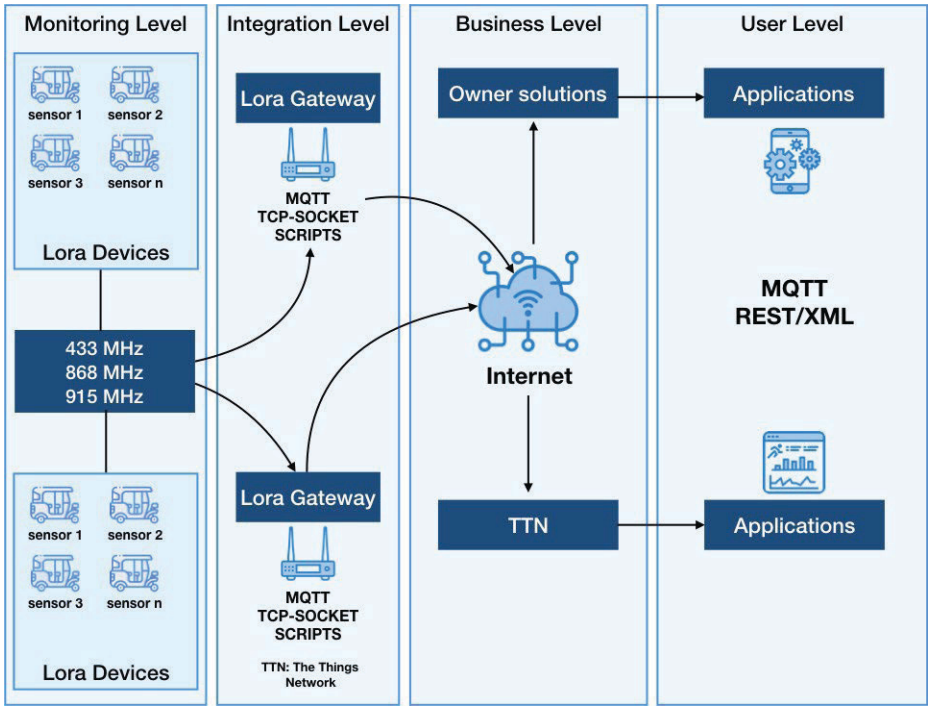


Fig.1. Proposal architecture for the rickshaw management fleet system

4.2 INTEGRATION LEVEL

Gateways are devices connected to the network server via a standard IP connection and it is capable of act as a transparent bridge, simply converting RF packets to IP packets and vice versa. The wireless communication takes advantage of the Long Range characteristics of the LoRa physical device, allowing a single-hop link between the end-device and one or many gateways. Even though gateways can operate using several different approaches for instance ISO 20992 (using the MQTT protocol), an owner script, or a TCP socket, the proposed architecture uses the MQTT protocol in order to make easy the subscription of systems to VTU events.

4.3 BUSINESS LEVEL

The business level of the LORA network-architecture, presented in this work, uses a redundant scheme for the connection, storage and access of information. On the one hand, the intermediate devices (gateways) are connected to the TTN network, through the internet, so that it is possible to create our own network and define our own service-rules under the TTN environment, taking advantage of a non-proprietary open network. On the other hand, the information collected by the gateways is, in turn, sent to a private network, where different technologies and services have been considered, not only to receive and store the information but also to perform all the analysis and business model. From this point of view, it is possible to provide information not only to the end user but also to all actors interested in the statistical analysis of data.

At the business level, it delivers those functions of the monitoring center. It is predicted that the monitoring center is connected to the Lora network and the traffic center of locality. So, bicycle flow produced on

roads can be controlled as well as keep in touch with several critical centers, such as emergency centers, hospitals, and police departments. According to these arguments, the business level is made up of a large number of technological elements. The following are the mininum requirements: a database system, a storage system, an application server. All of these components generate a large quantity of information that is used for providing several different ITS services.

4.4 USER LEVEL

This level includes the ITS technologies through which the ITS users will be able to consume ITS services. In this way, several different technologies that fulfill user requirements have been divided into the final user systems, the emerging corporate systems and the systems that deploy visual information directly in the infrastructure.

The first category includes systems such as PDAs, smart phones, laptops, navigators, and notebooks. For the second category, it is important to emphasize that in the near future, customers (pedestrians, travelers, car drivers, and ITS entities) will embrace retail techniques. Both of these techniques support and enhance customer and organizational relationships to provide better ITS valueadded services.
The users of this level are the many entities that can act as ITS users, such as hospitals, police, roadside emergency assistance centers, car drivers, pedestrians, and travelers, among other ITS users.

5. CONCLUSIONS

This proposal presents a rickshaw management fleet system that is designed as a tracking system and to deliver a detailed analysis of drivers' behavior. The presented model seeks tackles different variables that are not covered by other existing solutions and that are paramount to increase user security when using the service.
Under the proposal, each rickshaw will have an On-board Unit that uses a low-power wide-area network device to send information to the gateway only when is required, allowing to reduce energy consumption by the end-nodes.
There are various work directions from this point. First, it is required to integrate image recognition algorithms for traffic light violation and wrong lane detection. Second, a pilot test is being prepared to deploy a series of devices in the city with the end of monitoring the actual rickshaw service and provide a better service to the community while allows the head authorities to control the behavior of each vehicle.

REFERENCES

1. Understanding on-road practices of electric bike riders: An observational study in a developed city of china. Accident Analysis Prevention 59, 319 –326 (2013)
2. Cervero, R.: Transport infrastructure and the environment: Sustainable mobility and urbanism. IURD, Institute of Urban and Regional Development, University of California (2013)
3. Cervero, R., Golub, A.: Informal transport: A global perspective. Transport Policy 14(6), 445 – 457 (2007). https://doi.org/10.1016/j.tranpol.2007.04.011
4. CHERRY, C., HE, M.: Alternative methods of measuring operating speed of electric and traditional bikes in china-implications for travel demand
5. models. Journal of the Eastern Asia Society for Transportation Studies 8, 1424–1436 (2010). https://doi.org/10.11175/easts.8.1424
6. Dill, J., Gliebe, J.: Understanding and measuring bicycling behavior: A focus on travel time and route choice (2008)

7. Fan, Y., Chen, Q., Liao, C.F., Douma, F.: Smartphone-based travel experience sampling and behavior intervention among young adults (2012)
8. Johnson, M., Newstead, S., Charlton, J., Oxley, J.: Riding through red lights: The rate, characteristics and risk factors of non-compliant urban
9. commuter cyclists. Accident Analysis Prevention 43(1), 323 – 328 (2011). https://doi.org/10.1016/j.aap.2010.08.030
10. Kiefer, C., Behrendt, F.: Smart e-bike monitoring system: real-time open source and open hardware gps assistance and sensor data for electricallyassisted bicycles. IET Intelligent Transport Systems 10(2), 79–88 (2016). https://doi.org/10.1049/iet-its.2014.0251
11. Lee, C., Chan, L., Yang, C., Lee, G., Ciou, C.: The design and implementation of the e-bike physiological monitoring prototype system for cyclists. In: IEEE iWEM2011. pp. 161–165 (Aug 2011).
12. https://doi.org/10.1109/iWEM.2011.6021452
13. Lin, S., He, M., Tan, Y., He, M.: Comparison study on operating speeds of electric bicycles and bicycles: Experience from field investigation in
14. kunming, china. Transportation Research Record 2048(1), 52–59 (2008). https://doi.org/10.3141/204807
15. McLoughlin, I.V., Narendra, I.K., Koh, L.H., Nguyen, Q.H., Seshadri, B., Zeng, W., Yao, C.: Campus mobility for the future: the electric bicycle. Journal of
16. Transportation Technologies 2(01), 1 (2012)
17. Mekki, K., Bajic, E., Chaxel, F., Meyer, F.: A comparative study of lpwan technologies for large-scale iot deployment. ICT Express 5(1), 1 – 7 (2019). https://doi.org/https://doi.org/10.1016/j.icte.2017.12.005
18. Paefgen, J., Michahelles, F.: Inferring usage characteristics of electric bicycles from position information. p. 5 (01 2010). https://doi.org/10.1145/1899662.1899667
19. Prist, M., Freddi, A., Longhi, S., Monteriu`, A., Antonini, P.: Wireless sensor network based management system for electric bicycle-sharing. In: 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC).
20. pp. 1–6. IEEE (2016)
21. Prist, M., Freddi, A., Longhi, S., Monteriu`, A., Antonini, P.: Wireless sensor network based management system for electric bicycle-sharing. In: 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC). pp. 1–6 (June 2016). https://doi.org/10.1109/EEEIC.2016.7555741
22. Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D., Srivastava, M.: Biketastic: Sensing and mapping for better biking
23. Savage, N.: Cycling through data. Commun. ACM 53, 16–17 (09 2010). https://doi.org/10.1145/1810891.1810898
24. Talamini, G., Ferreira, D.P.: An informal transportation as a feeder of the rapid transit system. spatial analysis of the e-bike taxi service in shenzhen,
25. china. Transportation Research Interdisciplinary Perspectives 1, 100002 (2019). https://doi.org/10.1016/j.trip.2019.100002
26. Wu, C., Yao, L., Zhang, K.: The red-light running behavior of electric bike riders and cyclists at urban intersections in china: An observational study. Accident Analysis Prevention 49, 186 – 192 (2012). https://doi.org/10.1016/j.aap.2011.06.001, pTW + Cognitive impairment and Driving Safety
27. Yang, J., Hu, Y., Du, W., Powis, B., Ozanne-Smith, J., Liao, Y., Li, N., Wu, M.: Unsafe riding practice among electric bikers in suzhou, china: an observational study. BMJ Open 4(1) (2014). https://doi.org/10.1136/bmjopen-2013-003902

COST COMPARISON OF LAMBDA ARCHITECTURE IMPLEMENTATIONS USING PUBLIC CLOUD SOFTWARE AS A SERVICE

PEDRO F. P´EREZ-ARTEAGA1,2[0000–0001–8717–0595] CRISTIAN C. CASTELLANOS1,3[0000–0002–3301–8720] HAROLD CASTRO1,3[0000–0002–7586–9419]

YVES DENNEULIN4[]

LUIS A. GUZMAN1,2[0000–0002–6487–7579]

1 UNIVERSIDAD DE LOS ANDES, SCHOOL OF ENGINEERING, BOGOTA´, COLOMBIA HTTP://WWW.UNIANDES.EDU.CO

{PPEREZ,CC.CASTELLANOS87,HCASTRO,LA.GUZMAN}@UNIANDES.EDU.CO

2 CIVIL AND ENVIRONMENTAL ENGINEERING DEPARTMENT

3 SYSTEMS AND COMPUTING ENGINEERING DEPARTMENT

4 ENSIMAG, INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE, GRENOBLE, FRANCE

{YVES.DENNEULIN}@GRENOBLE-INP.FR

ABSTRACT.

Lambda architecture has gained high relevance for big data analytics by offering mixed and coordinated data processing: real time processing for fast data streams, and batch processing for large work-loads with high latency. However, concrete implementations over cloud infrastructures and cost comparisons are still not being sufficiently an-alyzed.

This paper presents a cost comparison of Lambda architecture implementations using Software as a Service (SaaS) to support IT de-cision makers when streaming analytics solution must be implemented. To do that, a case study of transportation analytics is developed on three public cloud providers: Google Cloud Platform, Microsoft Azure, and Amazon Web Services Cloud.

The evaluation is carried out com-paring deployment, configuration, development and performance costs in a public transportation delay monitoring case study assessing various concurrency scenarios.

Keywords: Lambda architecture · cost comparison · performance eval-uation · transport analytics · bus delay prediction · Software as a Service.

1. INTRODUCTION

Big data analytics (BDA) in real-time can provide up-to-the-minute insights to enterprise users, so that faster and better business decisions can be made. BDA requires the collecting of huge amount of data produced by multiple sources at high speed and process it with low latency using analytic algorithms.

In this context, Lambda architecture [8] has gained high relevance for BDA by offering mixed and coordinated data processing: real time processing for fast data streams, and batch processing for large workloads with high latency.

The Lambda architecture combines batch precomputed views and low-latency responses by building a series of layers which satisfy a subset of concerns. The batch layer stores a copy of the master dataset and precomputes the batch views. The batch layer stores an immutable, constantly growing, dataset, and computes arbitrary functions over the whole data-set to generate the batch views. This heavy workload implies high latency processing, and therefore the next layers compensate for this limitation.

The speed layer compensate the high latency of batch layer by precomputing the delta of data not processed by batch layer. The goal is to guarantee new data are included as soon as needed for the user queries thus offering speed views. The serving layer is a specialized distributed database that enables random reads on batch views. When new batch views are generated, the serving layer automatically swaps those in so that more up-to-date results can be queried.

Cloud computing is a enabler for big data solutions because it offers in-frastructure, storage, and processing capabilities that can be leased via pay-as-you-go models. These capabilities can be offered in different delivery models which are built one upon the other. Infrastructure-as-a-Service (IaaS) provides a self-contained environment comprised of IT infrastructure resources. Platform-as-a-Service (PaaS) offers a pre-configured cloud environment ready for the de-velopment and deployment of applications. Software-as-a-Service (SaaS) enables customers to use a high level functional services without incurring in license acquisition or software maintenance. This latter delivery model is oriented to decrease Total Cost of Ownership and increase Return On Investment.

Previous studies have proposed concrete implementations of Lambda archi-tecture [12, 5, 1] including cloud services [9, 7, 3]. However, concrete implementa-tions of Lambda architecture over SaaS and cost comparison have still not been sufficiently analyzed Cloud services facilitate the provisioning of near–infinite and elastic resources necessary for storing, and processing stream data analyt-ics and heavy batch workloads. For this reason, the public cloud is a natural environment to implement BDA solutions.

This paper presents a cost comparison of Lambda architecture implementa-tions taking benefit from SaaS delivery model to support IT decision making when streaming analytics solution must be implemented. To do that, a case study of transportation analytics is developed on three public cloud providers: Google Cloud Platform, Microsoft Azure, and Amazon Web Services (AWS) Cloud. The evaluation is carried out comparing deployment, configuration, de-velopment and performance costs in a public transportation delay monitoring case study assessing various concurrency scenarios.

This paper is organized as follows: Section 2 shows previous studies with implementations of Lambda architecture. Section 3 introduces the case study of transportation analytics. Section 4 describes the different implementations of Lambda architecture using SaaS. Section 5 summarizes the testing methodology.

Section 6 reports the results obtained. Section 7 tackles the discussion about such results. Finally, Section 8 outlines the conclusions.

2. RELATED WORK

The following previous works have been focused on implementations and opti-mizations of Lambda architectures deployed on IaaS and PaaS, but they neither tackle implementations on SaaS of different public vendors nor offer multi-factor cost comparison to support decision-making when a Lambda architecture solu-tion is instantiated. Pham in [9] proposes a flexibly adaptive cloud-based frame-work for BDA as a Service (BDAaaS) by implementing Lambda architecture for real-time analytics. The framework collects and analyzes data implementing concrete technologies for each Lambda layers.

These layers are deployed auto-matically over public cloud providers. Kiran et al. [7] present an implementation of Lambda architecture to construct a data processing on Amazon EC2 deliv-ered as a service for minimizing the cost of maintenance.

Thota et al. [10] present an architecture for integration to offers capabilities such as streaming and bulk processing and data services for cloud deployment. Grulich and Zukunft [4] pro-pose a streaming processing architecture for car information systems, and they validate scalability metrics on cloud infrastructure deployment. Similarly to pre-vious works, Dissanayake and Jayasena [2] offers a implementations of Lambda architecture for IoT analytics using AWS PaaS to address scalability, availability and performance quality attributes.

On the other hand, Gribaudo, Iacono and Kiran present in [3] a modeling ap-proach, based on multiformalism and multisolution techniques, for performance assessment of Lambda architecture implementations to optimize architecture de-signs. This work provides a user domain language approach to model a evaluate performance indices of Lambda architecture implementations regarding specific infrastructure, data speed and computation parameters, but they neither tackle software development effort nor cloud service costs regarding SaaS options pro-vided by different vendors.

3. CASE STUDY

Travel information service deal with the provision of static and dynamic informa-tion about the road transport network prior to and during the trips [6]. We are going to address a case study related to this service domain: real-time transport status information. Specifically, we use a service to provide the delay of trips within transportation system.

This information is generally provided by the ITS authority in real-time or near real-time to offer timely and accurate information to transport user. Delay monitoring in public transportation services requires combining the processing of large datasets of vehicle location and low latency to report the delay times to users in near real-time. This makes the delay monitor-ing service a typical use case to develop a big data solution applying Lambda architecture.

Our case study presents a proposed bus arrival time prediction with Lambda architecture. The developed architecture covers the batch layer using historical data with one day execution window, and the speed layer uses real-time data with five minutes execution window continually during the day.

The algorithm in both layers is an expected average delay in five-minute windows. These win-dows are

generated for each key composed of route id, stop id and window time. Additionally, delay average is grouped by day of the week. The window time is defined by the groups of trip updates reported within five minutes.

We take the Metro Vancouver's regional transportation (Translink) GTFS data set which is publicly available (Translink Open API 5) and real-time Trip Update data (GTFS real-time Open API) which provides real-time Vancouver's transportation data for the analysis.

3.1 TRANSLINK DATA SET

The open API of Translink serves trip updates data in GTFS realtime (proto-buffer format), and we send requests to collect feeds every sixty seconds. These data are collected during one week from December 11, 2017 to December 17, 2017, and 16 hours every day.

The GTFS real-time data contains just over 6,720 trip updates with 4,631,075 protobuffer files which are deserialized to JSON for-mat. In summary, these JSON files comprises 211 routes and 8,447 stops, each pair with a delay time to the next stop. The size of the data set (binary format) is 383 MB in 6,720 individual files. Each trip update contains the information presented in the appendix section at the end of the article.

3.2 STEPS NEEDED TO CALCULATE THE WAITING TIME

A Trip Update provides information in real-time about the trips in operation in the city of Vancouver. This means that the first step is to join the planned trips file in GTFS file with each Trip Update in GTFS real-time. This step is necessary in both layers.

In the next step, the speed layer receives every sixty seconds a Travel Update with approximately 45,000 JSON updates. The algorithm makes groups every five minutes (time window) with exactly five JSON updates. Then the speed layer assembles tuples with route id, stop id, and its expected delay average. At the end, the speed layer write each five minutes the results composed by stop id, route id, week day, time window and avg delay in the serving layer. Consequently, the preprocessed view with real-time information calculation is ready to respond users requests. The goal is to guarantee new data available as soon as needed for the user queries thus offering real-time views.

Simultaneously, the batch layer job is executed at the end of the day to compute whole stored raw data generating the same output (stop id, route id, week day, time window and avg delay). The batch layer writes each day the results over serving layer recomputing cumulatively historic data. This heavy workload implies high latency processing, and therefore the speed layer compen-sates this limitation.

Lastly, we implement and evaluate the lambda architecture using SaaS with a realistic and exhaustive tests described in the next Sections.

4. IMPLEMENTATION

To implement a Lambda architecture solution aligned to our case study we define architectural mechanisms for each layer. The ingestion process is imple-mented by means of event data transfer mechanism. the batch layer requires a batch processing engine combined with a resilient distributed file system to store the immutable master dataset.

The speed layer requires a streaming processing engine of low latency. Finally, the serving layer can be instantiated through rela-tional or column-family database regarding the model structure and offering low

latency. To compare each BDA SaaS, we implement versions for each Lambda layer and cloud platform regarding the architectural decisions and the SaaS cat-alog of each cloud vendor (Amazon, Google and Azure). In each layer of Lambda architecture, we select the service with the highest level of abstraction, serverless delivery model and the best Service Level Agreements (SLA) in terms of avail-ability and performance. This selection is made for two main reasons: to avoid low level implementation and to make the metrics comparable.

4.1 AWS IMPLEMENTATION

The AWS implementation is depicted in Figure 1. Speed layer uses Kinesis Data Streams to ingest GTFS messages and to send them to Kinesis Analytics to be processed in real-time. The processing output (speed views) are stored into S3 batch bucket using an AWS Lambda function.

In batch layer, Kinesis Firehose ingests the raw data and stores it into a S3 bucket. Raw data is read and pro-cessed by an AWS Glue job to be persisted as batch views in S3 result bucket. The serving layer uses Amazon Athena to directly perform queries in standard SQL over speed and batch views stored in S3 buckets.

4.2 GOOGLE CLOUD IMPLEMENTATION

The Google Cloud implementation employs Dataflow service in both speed and batch layer and its detail is presented in Figure 2. The speed layer ingestion is developed by means of a topic in Cloud Pub/Sub which passes the GTFS messages to a Dataflow speed job.

This job aggregates the calculations and stores them into Google Cloud BigQuery. In the batch layer, Pub/Sub service persists messages into Cloud Datastore as raw data. Then, a batch Dataflow job reads the raw data and aggregates delay averages to write the batch views into BigQuery. BigQuery is the Serving layer SaaS used to persist and query the views using a SQL-like scripts.

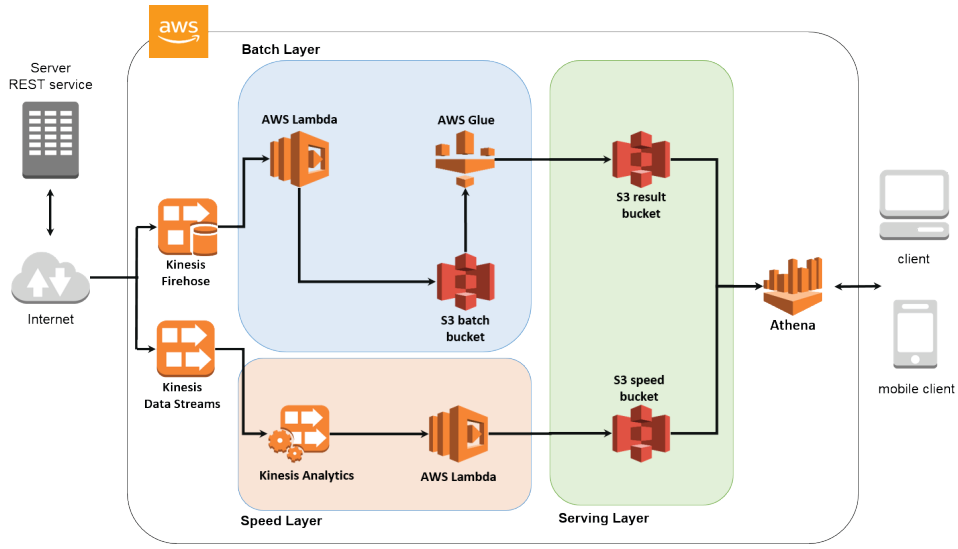


Fig. 1. Implementation in AWS

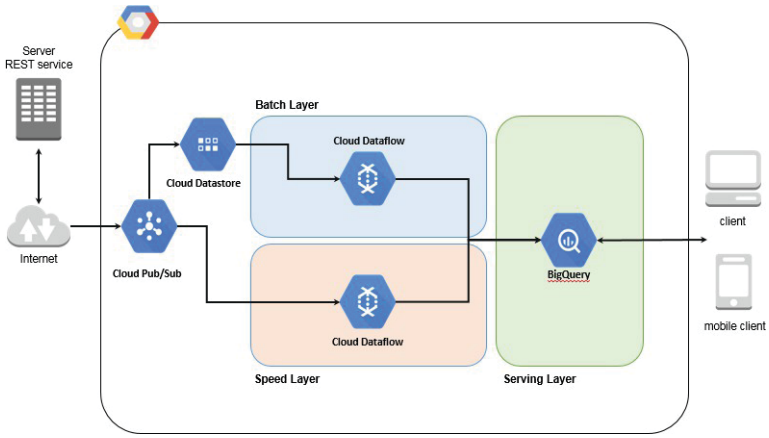


Fig. 2. Implementation in Google Cloud

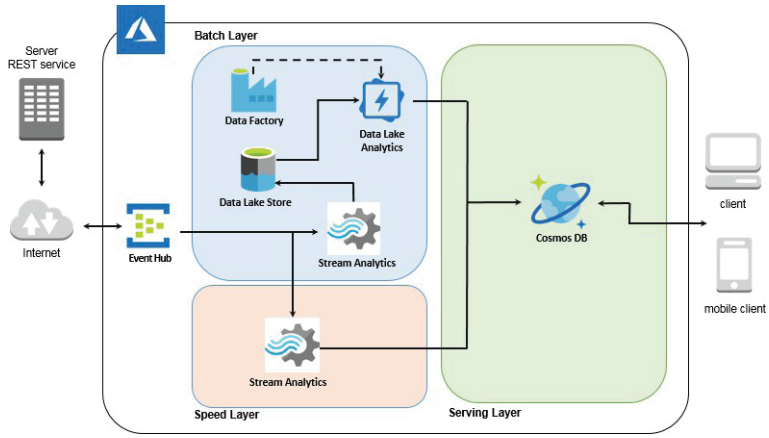


Fig. 3. Implementation in Microsoft Azure

4.3 AZURE IMPLEMENTATION

For Azure implementation, represented in Figure 3, speed layer uses EventHub to ingest GTFS messages and Stream Analytics service processes them in real-time. The processed speed views are stored in Cosmos DB. In the batch layer, raw data is persisted into Data Lake Store using a Stream Analytics job. The raw data is read by a Data Lake Analytics job which is scheduled through Data Factory. The Data Lake Analytics job makes the calculations and stores the results in Cosmos DB. The serving layer is built as a Cosmos DB service which stores the batch and speed views and it offers a SQL-like interface.

5. TESTS

We evaluate the three implementations of the Lambda architecture presented in Section 4 to compare performance, development/configuration efforts, and service costs in each layer using GTFS trip updates data set introduced in Section

3.1. Table 1 summarizes the metrics evaluated for each layer. The metrics used to compare the cost of the implementations are calculated by layer with the purpose of architects, administrators and developers can evaluate

and select the best SaaS candidate for each layer regarding performance requirements, time to market and budget.

5.1 PERFORMANCE TEST

To compare performance for each public cloud provider and layer, we define metrics related to reading time, processing time, writing time, response time, and response time vs active threads. In the speed layer, we measure processing time for each microbatch to evaluate the processing speed offered. In the batch

Table 1. Comparison metrics for layer

Metric		Layer		
		Speed	Batch	Serving
Performance	Reading time		X	
	Processing time	X	X	
	Writing time		X	
	Response Time			X
	Time vs Threads			X
Development/Configuration Effort		X	X	X
Service Costs		X	X	X

layer, we collect reading time of raw data, processing time and results writing for each daily execution. In the serving layer, we take response time and response time vs thread metrics using stress test with a ramping-up depicted in Figure 4 to evaluate the final user experience when the delay service is consumed.

The experiment involves a simulation of the consumption of GTFS dataset accelerated up to 60 times, what implies consuming a GTFS feed each second. At the same time, serving layer is assessed by automated stress test implemented in JMeter which launches JDBC queries simulating delay service requests made by the users. The request's ramp-up reflects a real demand scenario depicted in [11], where there are time slots of low, medium and high demand during the day.

Hence, the Figure 4 details the number of requests per day (one day = sixteen minutes in 60X simulation). The whole simulation (seven days) in each platform takes 112 minutes where batch job execution is performed each 16 minutes, and speed job each 5 seconds.

5.2 DEVELOPMENT AND CONFIGURATION

Regarding the development and configuration effort quantification, we track the time invested by each programmer to develop each layer. To have a comparable effort metric, we ensure that developers have similar technical skills. The de-velopment tasks include training, coding and testing.

Thus, trip update JSON parsing, join, filter and aggregate operations in each layer (speed and batch) are registered in hours as ETL development. Time invested in scripts building for serving layer (in most cases SQL-like) are also recorded. Additionally, SaaS con-figuration tasks such as scheduling, parameters setting and service provisioning are also timed.

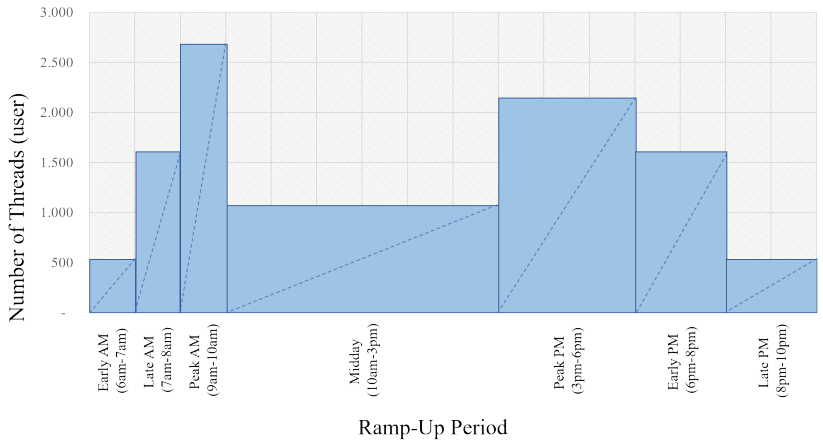


Fig. 4. Number of threads per Time slots (Ramp-up)

5.3 SERVICES COST

Due to different SaaS princing models for each layer, the economic cost can be calculated according the demand of the tasks, requests, processing, storage or resources.

So, we sum these costs up to obtain a cumulative cost reported by vendor's billing service for each layer. The total cost of the experiment (seven days) is projected to a monthly fee.

6 RESULTS

The case study allows us to evaluate the performance, development effort and cost of each public cloud. The results of this evaluation are described in this section.

6.1 PERFORMANCE

The performance test of the batch layer involves the cumulative processing of trip update files each day. Approximately, each day were collected 1000 trip update files comprises by 700.000 JSONs. In total, 6.720 files and 4.631.075 JSONs were collected to be processed.

Before starting the processing in the batch layer, the raw of Trip Updates should be read, for this reason Figure 5 presents the average reading time for each implementation.

The average reading time of AWS Glue in AWS S3 storage is the most stable and efficient, while the other batch services take 12 times (Google Cloud) and 18 times (Azure) more time reading raw data. The average reading time of the Cloud Datastore service in Google Cloud has a constant increase as the number of Trips Update increases every day. And finally, the average readingtime of the Data Lake Store service in Azure has the highest increase until the fifth day, after that day the average reading time has a decrease, which may reflect a scaling of the service.

Days	Number of Trip update files	Google Cloud	AWS	AZURE
1	1,177	15	4.1	33
2	2,161	26	4.2	63
3	3,114	41	4.2	86
4	4,150	55	4.4	102
5	5,064	60	4.1	133
6	6,048	70	4.3	79
7	6,720	95	4.3	57
Total	28,434	362	30	553

Fig. 5. Average reading time in seconds to Batch layer

After reading the files the next step is to calculate the waiting time described in section 3.2. This processing time is shown in Figure 6. The AWS Glue service that does the processing of the batch layer in AWS, again is the most consistent and efficient, since the processing time is almost constant below two seconds in each execution, despite the increasing amount of files. In contrast, the Google Cloud Dataflow service has the lowest processing performance with peaks almost four seconds, twice the processing time of AWS. Data Lake Analytics in Azure is the most sensitive to the number of processed files, and similar to reading time the service seems to have scaled during the fifth and sixth day.

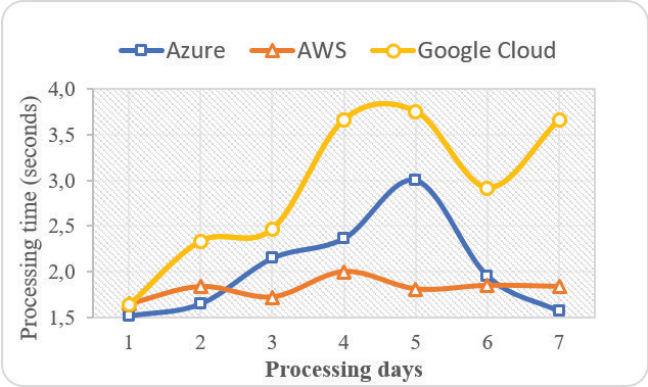


Fig. 6. Average processing time for Batch layer

The final step of batch processing is to write in the serving layer. The average writing time is shown in Figure 7. Amazon S3 service continue with consistent be-havior offering the best performance. Conversely, Google BigQuery presents the worst average writing times describing a decreasing trend. In addition, Cosmos DB service presents intermediate average writing times with a slight degradation observed in the last two days.

Days	Number of records	Google Cloud	AWS	AZURE
1	571,917	88	54	62
2	1,024,509	102	39	91
3	1,460,840	145	47	111
4	1,934,729	196	49	68
5	2,303,636	245	46	118
6	2,642,347	326	40	232
7	2,860,524	380	42	239
Total	12,798,502	1482	318	921

Fig. 7. Average writing time in Batch layer

The processing times obtained in speed layer are constant in all platforms constrained to real-time windows, and for this reason we do not consider valuable to compare them.

The metric of serving layer performance in respect of response time is shown in Figure 8. It is worthy of note that at beginning of the stress test, all ser-vices start with the highest latency specially noticeable in Google serving layer, but when test moves forward, the latency is reduced. Cosmos DB depicts the lowest average response times, followed by AWS Athena and Google BigQuery respectively.

6.2 DEVELOPMENT AND CONFIGURATION

The effort required for learning, development, configuration and deployment was measured for each developer. Figure 2 shows that the total of development hours is the highest for AWS, followed by Azure and Google respectively. Google Dataflow implementation requires the lowest development time in the whole im-plementation. Detailing the development effort in the speed layer, the greatest effort is required in AWS Kinesis service. Google Dataflow implementation seems to require the lowest development time probably due to its unified programming model. In the batch layer, the Data Lake Analytics service requires the greatest effort for its implementation, while the others cloud services report similar in-vestment of time. Finally, in the serving layer, the AWS Athena-S3 integration requires the greatest effort time, while Azure Cosmos DB demands the lowest development time.

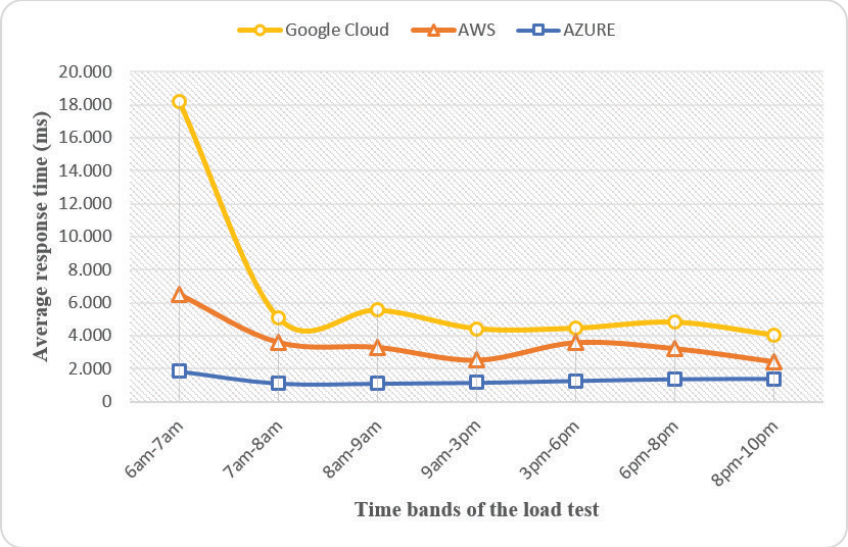


Fig. 8. Average response time for Serving layer

Table 2. Development time of Lambda architecture on each public cloud

	Google Cloud	AWS	Azure
Speed layer	26.1	42.8	37.4
Batch layer	31.6	31.5	39.7
Serving Layer	16.7	26.2	8.2
Total (hours)	74.4	100.5	85.3

6.3 SERVICES COST

Each implementation of the Lambda architecture is deployed in different public cloud providers, we define and calculate the costs required to replicate a similar case study, with data similar to Vancouver's transportation system and operate them for 4 weeks. As a result, Figure 9 presents a summary of the monthly fees generated by each provider during the simulation. The highest monthly cost is generated by Azure, and specifically due to the high cost of the Cosmos DB ser-vice. AWS Glue and Kinesis, both in batch and speed layer respectively, are the highest individual costs in these layers with respect to the other infrastructures. Google cloud is the least expensive provider in all layers and with remarkable difference. Finally, regarding the learning curve, Google Cloud free tier enables an inexpensive proof of concept with these SaaS compared to other vendors free tier.

	Google Cloud	AWS	AZURE
Speed layer	$ 32.64	$ 115.40	$ 64.24
Batch layer	$ 12.04	$ 112.60	$ 45.38
Serving layer	$ 13.33	$ 15.08	$ 159.41
Other services	$ 3.20	$ 0.56	$ 10.99
Monthly cost (USD)	$ 61.21	$ 243.64	$ 280.02

Fig. 9. Infrastructure monthly costs (USD)

7. DISCUSSION

This document presents a comparison of costs in development, and deployment the same case study over Lambda architecture using three different public cloud providers (Google Cloud, Microsoft Azure, and Amazon Web Services) with the main goal of identifying how different public cloud providers providers with the same architecture deployment can affect the infrastructure cost of running and performance with concurrence users.So as to get valid results, we implementing three version of the Lambda architecture and deploy each one using a different public cloud provider.

As a result of the developing and testing process of the three implementation deployed, we could understand the challenges that must be overcome to use the Lambda architecture.

8. CONCLUSIONS

This work presented lambda architecture implementations in different public cloud vendors. Also, this research offered a comparison of such implementations to support decision makers when they need to select specific-vendor's SaaS in the context of BDA. Based on the results obtained, we recommend the most suitable SaaS for each layer depending on the criteria selected.

In terms of performance, AWS obtained the best metrics in batch and speed layer. In batch layer, AWS reported the best performance in reading, processing and writing time, whereas Google Cloud seems to be affected by increasing data size. Focusing on serving layer performance, Azure presented a constant and efficient behavior compared to other competitors.

Regarding the time-to-market, AWS required more man-hours, specially in speed and serving layer. Azure had the fastest development in serving layer, but batch layer implementations required more effort because it implied Data Lake Store, Stream Analytics, Data Factory and Data Lake Analytics services development and integration. Google Cloud development was the fastest, which could be caused by unified programming model for batch and speed processing offered by Google Dataflow.

In terms of cost services, Azure was the most expensive provider in serving layer, whereas AWS consumed more credits in Serving Layer due to Cosmos DB service. In contrast, Google Cloud presented the lowest price in all layers and it offers the widest free tier to initiate the training.

In summary, when performance is a strong concern, despite the high cost, AWS (in batch and speed layer) is the best choice and Azure (in serving layer) should be selected to obtain the best response times. If the time-to-market guides the SaaS selection, Google Cloud is recommended although the performance could be affected. Finally, if service pricing is an important constraint, Google Cloud again offers the best choice with a factor of 1/4.

ACKNOWLEDGEMENTS

This research was carried out by the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA), supported by the Ministry of In-formation Technologies and Telecommunications of the Republic of Colombia (MinTIC) through the Colombian Administrative Department of Science, Tech-nology and Innovation (COLCIENCIAS) within contract No. FP44842-anexo 46-2015. A special thanks to CAOBA's members: Miguel Barrera, Felipe Gonzalez-Casabianca, Miguel Rodriguez and Camilo Ortiz.

REFERENCES

1. Batyuk, A., Voityshyn, V.: Apache storm based on topology for real-time processing of streaming data from social networks. In: 2016 IEEE DSMP.
2. 345–349. IEEE (aug 2016). https://doi.org/10.1109/DSMP.2016.7583573,
3. http://ieeexplore.ieee.org/document/7583573/
4. Dissanayake, D.M.C., Jayasena, K.P.N.: A cloud platform for big iot data analytics by combining batch and stream processing technologies. In: 2017 NITC. pp. 40–45 (Sept 2017). https://doi.org/10.1109/NITC.2017.8285647
5. Gribaudo, M., Iacono, M., Kiran, M.: A performance modeling frame-work for lambda architecture based applications. Future Generation Com-puter Systems(jul 2017). https://doi.org/10.1016/j.future.2017.07.033, http://www.sciencedirect.com/science/article/pii/S0167739X17315364 http://linkinghub.elsevier.com/retrieve/pii/S0167739X17315364
6. Grulich, P.M., Zukunft, O.: Smart stream-based car information systems that scale: An experimental evaluation. In: 2017 IEEE iThings. pp. 1030–1037 (June 2017). https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.181
7. Hasani, Z., Kon-Popovska, M., Velinov, G.: Lambda architecture for real time big data analytic. ICT Innovations pp. 133–143(2014), http://proceedings.ictinnovations.org/attachment/paper/299/lambda-architecture-for-real-time-big-data-analytic.pdf
8. ISO: Intelligent transport systems - Reference model architecture(s) for de ITS sector. Part 1: ITS service domains, service groups and services (2001), www.iso.org
9. Kiran, M., Murphy, P., Monga, I., Dugan, J., Baveja, S.S.: Lambda ar-chitecture for cost-effective batch and speed big data processing. In: 2015 IEEE International Conference on Big Data (Big Data). pp. 2785–2792. IEEE (oct 2015). https://doi.org/10.1109/BigData.2015.7364082, http://ieeexplore.ieee.org/document/7364082/

10. Marz, N., Warren, J.: Big Data, Principles and best practices of scalable real-time data systems. Manning Publications Co. (2015), https://www.manning.com/books/big-data
11. Pham, L.M.: A Big Data Analytics Framework for IoT Applications in the Cloud. VNU Journal of Science: Computer Science and Communication Engineering 31(2), 44–55 (2015)
12. Thota, C., Manogaran, G., Lopez, D., Sundarasekar, R.: Architecture for Big Data Storage in Different Cloud Deployment Models. In: Hand-book of Research on Big Data Storage and Visualization Techniques,
13. 196–226. IGI Global (2018). https://doi.org/10.4018/978-1-5225-3142-5.ch008, http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-5225-3142-5.ch008
14. TransLink: 2011 Metro Vancouver Regional Trip Diary Survey Analysis Re-port. Tech. rep., TransLink, Vancouver (2013), https://www.translink.ca/ /me-dia/documents/about translink/media/2012/2011 metro van trip diary survey briefing 1.ashx
15. Villari, M., Celesti, A., Fazio, M., Puliafito, A.: AllJoyn Lambda: An architecture for the management of smart environments in IoT. In: 2014 International Conference on Smart Computing Workshops. pp. 9–
16. 14. IEEE (nov 2014). https://doi.org/10.1109/SMARTCOMP-W.2014.7046676,
17. http://ieeexplore.ieee.org/document/7046676/

EVENT DETECTION IN COLOMBIAN SECURITY TWITTER NEWS USING FINE-GRAINED LATENT TOPIC ANALYSIS

VLADIMIR VARGAS-CALDER´ON1[0000–0001–5476–3300],

NICOLAS PARRA-A.1[0000–0002–1829–4399], JORGE E. CAMARGO2[0000–0002–3562–4441],

AND HERBERT VINCK-POSADA1

1 GRUPO DE SUPERCONDUCTIVIDAD Y NANOTECNOLOGÍA, DEPARTAMENTO DE FÍSICA, UNIVERSIDAD NACIONAL DE COLOMBIA, 111321, COLOMBIA

{VVARGASC,NPARRAA,HVINCKP}@UNAL.EDU.CO

2 UNSECURELAB RESEARCH GROUP, DEPARTAMENTO DE INGENIERÍA DE SISTEMAS E INDUSTRIAL, UNIVERSIDAD NACIONAL DE COLOMBIA, 111321, COLOMBIA

JECAMARGOM@UNAL.EDU.CO

ABSTRACT.

Cultural and social dynamics are important concepts that must be understood in order to grasp what a community cares about. To that end, an excellent source of information on what occurs in a com-munity is the news, especially in recent years, when mass media giants use social networks to communicate and interact with their audience.

In this work, we use a method to discover latent topics in tweets from Colombian Twitter news accounts in order to identify the most prominent events in the country. We pay particular attention to security, violence and crime-related tweets because of the violent environment that surrounds Colombian society. The latent topic discovery method that we use builds vector representations of the tweets by using FastText and finds clusters of tweets through the K-means clustering algorithm.

The number of clusters is found by measuring the CV coherence for a range of number of topics of the Latent Dirichlet Allocation (LDA) model. We finally use Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction to visualize the tweets vectors. Once the clusters related to security, violence and crime are identified, we proceed to apply the same method within each cluster to perform a fine-grained analysis in which specific events mentioned in the news are grouped to-gether. Our method is able to discover event-specific sets of news, which is the baseline to perform an extensive analysis of how people engage in Twitter threads on the different types of news, with an emphasis on security, violence and crime-related tweets.

Keywords: Event detection · Latent topics · Fasttext · Security · Vio-lence · Crime

1. INTRODUCTION

Twitter3 is a social network that has been used worldwide as a means of news spreading. In fact, more than 85% of its users use Twitter to be updated with news, and do so on a daily basis [23]. Users behavior of this social network has been found efficient in electronic word-of-mouth processes [29], which is a key component for the quick spreading of breaking news. This would lead to think that news-related content occupies the majority of the tweets volume. However, on average, the proportion of news-related content to the total content of tweets is 1% worldwide, but increases dramatically (up to 15%) in countries in conflict [14]. An extrapolation of these findings indicates that Colombia might have a high content of news-related tweets, since it is well-known that Colombia is one of the most violent countries in the world, and has been for decades [10].

On the other hand, the virality or importance of a tweet conveying news-related information is a relevant measure of what is critical for a community. Therefore, the study of news spreading in a community gives a clear idea of citizens interactions around central topics of interest. Particularly, we are interested in examining how people react to news related to security, crime and violence because this would expose the mechanisms of collective reactions of rejection, acceptance, conflict, among others.

This has been considered in case-studies such as Ref. [18], where messages containing hate or violent speech were identified after Charlie Hebdo's famous shooting, allowing researchers to identify spatiotemporal and textual patterns in the produced tweets after the mentioned disruptive event. Other similar case-studies include the analysis of how people react to homicides in London [12], to polio health news [25], and to the after-math of violence on college campuses [11]. Also, social networks and technology have been signaled as tools used by young people to inflict violent acts against others [6,2,16].

On a more general ground, the study of these individual or collective reactions is a problem tackled by sentiment analysis, whose objective is to determine whether the sentiment contained in a text is positive or negative, and to what extent [24,1,19,20]. Applied to security-related content in social networks, sentiment analysis could be important in designing and implementing public policies regarding security, crime and violence, as well as educational campaigns where people are taught to communicate their opinions in a non-violent way. However, in order to achieve this, it is desirable to segment news-related tweets so that different topics can be differentiated from one another as we expect Twitter users to react quite differently depending on the security topic they react to. This field is known as topic discovery.

Several proposals for topic discovery are available, among which many La-tent Dirichlet Allocation variants and modifications are available. For instance, Ref. [30] presents an LDA-based model that relates the topic of a scientific paper with the content of the documents that it cites. The proposed algorithm allows to know the evolution of a research topic by measuring whether a topic is important (as seen by the scientific community) or not. Furthermore, Ref. [8] used LDA with variational Gibbs sampling to found general terms that associate the re-views of users in e-commerce web sites. This with the intention of improving the experience of new users. Moreover, we have recently combined word-embedding methods and K-means to discover topics and have good interpretability results [28,27].

Thus, in this paper we exemplify a method for topic discovery applied to Colombian news-related tweets that is accurate in the task of segmenting tweets. This method can work at different granularity levels depending on the corpus to be analyzed. In this case, we have a corpus of security-related tweets, so that the method will group tweets into the different sub-topics such as murders, robberies, among others. The workings of the method will be detailed in section section 2, and the main results will be presented in Section section 3. Finally, we provide some conclusions in Section section 4.

2. METHODS AND MATERIALS

In this section, we describe the dataset used in our research, as well as the methods to perform fine-grained latent topic analysis to process all the data. The method is largely based on our previous work [28]. Tweets from Colombian news Twitter account @NoticiasRCN were collected from 2014 to the present. A total of 258,848 tweets were published by @Notici-asRCN in this period. The method hereafter mentioned was applied in Ref. [28] to this large corpus at a coarse-grained scale to discover news topics, allowing us to pinpoint groups of tweets sharing semantic content. It was possible to detect tweets regarding to politics, sports, Colombian armed conflict, extreme violence, organised/common crime, among others. In this paper, we focus on the groups of extreme violence and organized/common crime, which contain a total of 47,229 tweets, accounting for an 18.2% of all published news. We excluded the Colombian armed conflict, as this is not normally connected to events occurring in cities.

We pre-processed these tweets related to security, crime and violence by removing punctuation, links, hashtags and mentions, we lowercased the text and performed lemmatisation with spaCy's adapted Spanish lemmatiser4.

As the objective of our work is to find a number of topics and its members that are helpful for further analysis, the first task to solve is to methodologically find the number of topics. Our proposal is to combine a topic modelling tool with a measure of how well this tool performed. Therefore, we trained a Latent Dirichlet Allocation (LDA) model [3] to soft-cluster tweets into topics and then used CV coherence [21,26] to measure the interpretability of the LDA results. What LDA does is to assign to documents probabilities to belong to different topics (an integer number k provided by the user), where these probabilities depend on the occurrence of words which are assumed to co-occur in documents belonging to the same topic.

This assumption is called a sparse Dirichlet prior. Thus, LDA exploits the fact that even if a word belongs to many topics, oc-curring in them with different probabilities, they co-occur with neighbouring words in each topic with other probabilities that help to better define the topics. The best number of topics is the number of topics that helps the most human interpretability of the topics. This means that if the topics given by LDA can be well-distinguished by humans, then the corresponding number of topics is acceptable. As mentioned before, a way of measuring this interpretability is the calculation of CV coherence, which, to the knowledge of the authors, is the measure with largest correlation to human interpretability. The optimum number of topics can be found at the maximum of CV (k).

Once the number of topics has been determined, we proceed to find vector-embedding representations of tweets, as they have been previously shown to yield superior results in topic modelling with respect to LDA [9]. Here, we use the word2vec-based [17,7,22] FastText model [4], which essentially uses sub-word information to enrich embeddings generated by a neural network that predicts neighbouring words. The job of FastText is to reduce the dimensionality of one-hot-encoded words (which may be in very large vector spaces of the size of the corpus vocabulary) to a low-dimension and dense vector space (of dimension N , selected by the user), where dimensions store highly correlated semantic relations between words and strings of characters. Here a tweet is represented by the sum of the individual vector representations of each word in the tweet.

In this low-dimensional vector space, K-means clustering [13] is performed for k clusters (i.e. the number of topics found with the LDA-CV coherence method) in order to hard-cluster the vector representations of the tweets. K-means is a common clustering technique that minimises within-cluster dispersion. Each cluster contains tweets belonging to the same topic.
In order to visualise the clusters and interpret their contents, we performed dimensionality reduction with the Uniform Manifold Approximation and Projection (UMAP) method [15] to plot vectors onto a

2 dimensional space. UMAP learns the topology of the data to be reduced in dimension by learning a projection of this data onto a lower-dimensional space where the projection preserves, as much as possible, the fuzzy topological structure of the manifold described by the vectors.

Additionally, the Python's Bokeh library [5] was used to create interactive plots of the UMAP reduced representation of FastText vectors, al-lowing us to quickly examine the structure of the clusters, as well as to read representative tweets of each cluster to determine and label their corresponding topics.

3. RESULTS AND DISCUSSION

Since LDA is a probabilistic method that starts its learning with some random parameters, we measured CV coherence 64 times for each number of topics k ranging from 2 to 59, and averaged the measurements. The results are shown in fig. 1. It is important to note that a maximum is not reached. However, a saturation of the CV coherence takes place, making it difficult to select a number of topics. By visual inspection and the use of the so-called elbow method, we pick two different numbers of topics (10 and 16) and analyse them separately in order to determine what the best number of topics is.

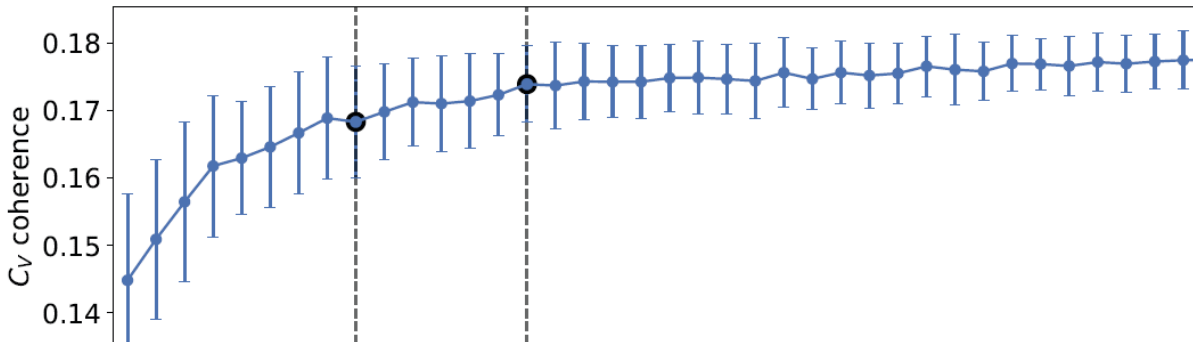


Fig. 1. CV coherence as a function of the number of topics. Error bars indicate one standard deviation error.

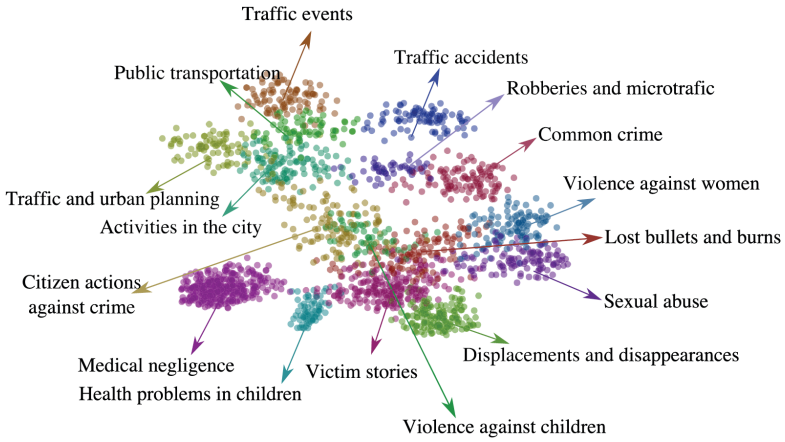


Fig. 2. K-means clustering results with 16 topics and their respective names.

The distribution of tweets along the 16 topics is shown in fig. 3, where news were slightly concentrated

on security-related sub-topics like activities in the city, citizen actions against crime, victim stories and common crime. The within-cluster dispersion is comparable between topics, implying that clusters are equally diffused.

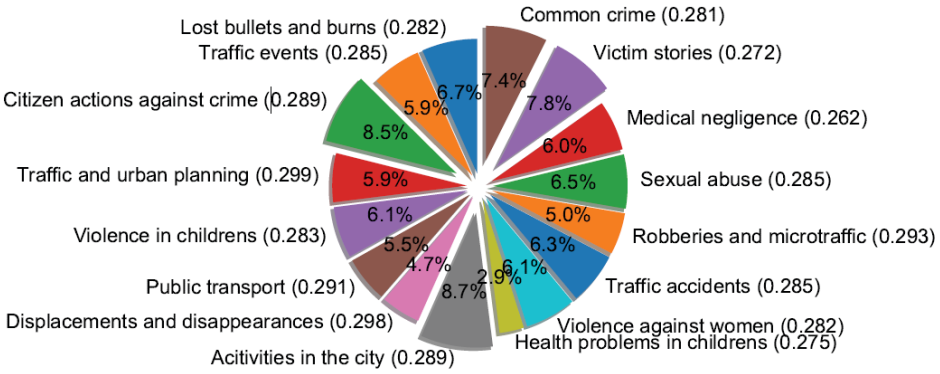


Fig. 3. Number of tweet distribution in the 16 topics, the number in parenthesis that follows each label is the within-cluster dispersion $N_{T_i}^{-1} \sum_{k \in T_i} d_k^i$, where Ti is the set of tweets corresponding to topic i, NTi is the number of tweets for that topic and d_k^i is the distance of the k-th tweet vector representation to the centroid of the i-th topic.

Regarding the case of only 10 clusters, in fig. 4 we plot the UMAP-reduced vectors that are most representative of each cluster, just as in fig. 2. We find that this number of topics is better than 16 clusters, since the identification of the topic was clearer in the case of 10 clusters when reading the 15 most representative tweets of each cluster.

This holds true even for three different topics referring to traffic, as they topics can be well-differentiated (note that UMAP plots those groups close one another because they all contain traffic-related tweets).

For instance in the traffic and urban planning cluster we find news such as Bogot´a's Secretaría de Movilidad talks about changes that users will find in public transportation this Monday which is not directly related with security, crime or violence topics.

On the other hand, in the traffic accidents cluster we find news such as In Cartagena, two bus drivers left their vehicles in the middle of the highway to solve their differences by fighting against each other.

Finally, in the events in the traffic cluster, an example of a news tweet is North highway is collapsed by a triple-crash. The air patroller recommends to take alternate routes.

These tweets exemplify the difference of the three traffic-related clusters. More-over, by comparing figs. 2 and 4, it is clear that some clusters are barely changed when increasing the number of topics from 10 to 16 because they are well-defined topics that can be interpreted easily from the human perspective.

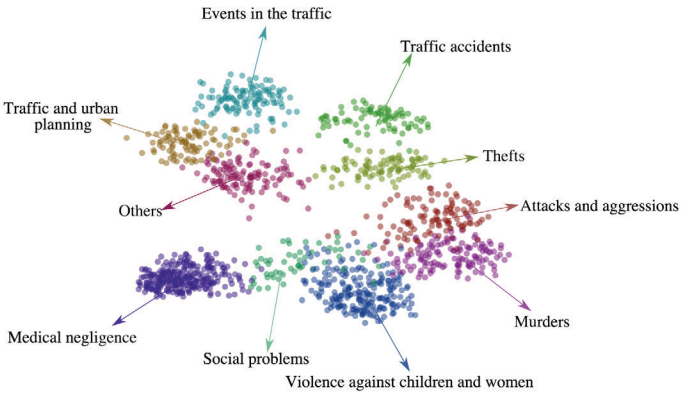


Fig. 4. Same as fig. 2 but for 10 topics.

Finally, the distribution of the tweets along the 10 different topics are shown in fig. 5. We found in this number of topics, that the clusters correspond to more general topics, allowing a better separation of the tweets. It is noteworthy to state that the "others" cluster mixes different security problems such as aggression against animals, government infringements to provide services to people, among others.

4. CONCLUSIONS

In this paper we presented the application of a latent topic discovery method at a fine-grained level to segment Colombian news published through Twitter in different sub-topics regarding security, crime and violence. We were able to find interpretable groups of tweets published by news-media giant @NoticiasRCN, where each group referred to different sorts of security, crime and violence issues.

We identified clear labels that summarize the content of the tweets belonging to each topic: attacks and aggressions, traffic and urban planning, thefts, traffic accidents, social problems, events in the traffic, violence against children and women, medical negligence, murders and others. An important application of

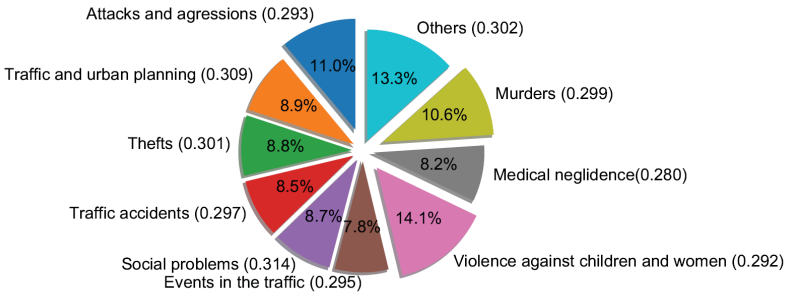


Fig. 5. Same as fig. 3 but for 10 topics.

the methodology presented in this paper is to detect violent events that go unreported to the police. Furthermore, the tools that were shown configure a critical channel for monitoring violent actions that attempt against the security of women, children, minorities and crime victims.
Moreover, our method allows the automatic classification of new security-related tweets.
Our method contributes to the segmentation of tweets to better address issues in each security front.

What we will develop in future work is the characterization of people's reaction to different types of security, crime and violence related issues, and to identify violent behavior in social networks, as this is a cornerstone to understanding social and cultural dynamics in our communities.

REFERENCES

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011). pp. 30–38 (2011)
2. Barlin´ska, J., Szuster, A., Winiewski, M.: Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and em-pathy. Journal of Community & Applied Social Psychology 23(1), 37–51 (2013). https://doi.org/10.1002/casp.2137, https://onlinelibrary.wiley.com/doi/abs/10.1002/casp.2137
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. arXiv e-prints arXiv:1607.04606 (Jul 2016)
5. Bokeh Development Team: Bokeh: Python library for interactive visualization (2018), https://bokeh.pydata.org/en/latest/
6. Draucker, C.B., Martsolf, D.S.: The role of electronic communication technology in adolescent dating violence. Journal of Child and Adolescent Psychiatric Nursing 23(3), 133–142 (2010)
7. Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al.'s negative- sampling word-embedding method. arXiv preprint arXiv:1402.3722 (2014)

OPEN SOURCE TOOL FOR VEHICULAR TRAFFIC SIMULATION AT VIRTUAL ENVIRONMENTS, STUDY CASE*

JUAN S.VARGAS MOLANO1[0000–0001–7881–1928], NICOLAS CASANOVA1[0000–0002–0841–6145], OSCAR CARRILLO2[0000–0001–5081–1774], AND WILSON J. SARMIENTO1[0000–0001–7903–8316]

1 MULTIMEDIA RESEARCH GROUP, UNIVERSIDAD MILITAR NUEVA GRANADA, COLOMBIA

{U1201559,U1201850,WILSON.SARMIENTO}@UNIMILITAR.EDU.CO.COM

2 CPE LYON/INSA LYON, FRANCE

OSCAR.CARRILLO@INSA-LYON.FR

ABSTRACT.

In last years end-user applications of virtual environments have been a high growth. A reason for this is the low cost of equipment and several alternatives to software development. In this way, virtual environments have penetrated multiples sectors of application, including planning and simulation of smart cities.

However, what is real alterna-tives to develop a smart cities solution with currents tools? This work presents a study case of a vehicular traffic simulator using open source tools. The virtual environments run over Linux and were developed in Unreal Engine 4. The study case shows the integration of open source tools.

For example, OpenStreetMap allows to obtain geodata, Carla to model autonomous vehicular actions in order to simulate complex traffic over a city, and AirSim to implement physical behavior of used vehicle to generate a believable immersive experience. The study case was im-plemented in a small sector of the north of Bogota´ city, which has dense vehicular traffic.

Keywords: Virtual Environments · Vehicular Traffic Simulator · Open Source Tool.

INTRODUCTION

Computer traffic simulation is a field that has been of growing interest in the research community in the past few years. It allows us to create accurate predic-tions of traffic behavior based on a diverse amount of data. Then, these predic-tions can create models that are used in a real environment for further testing.

For this purpose, different existing simulation packages differ between them with their software architecture and design paradigms used for traffic description. They are useful, but most of these are proprietary and not open to the public.

Consequently, there is very few open-source software for traffic simulation in virtual environments. Only recently have laboratories started building these open source frameworks for anyone to use. However, these new tools are still in it's infancy and have a lot of limitations compared to proprietary software.

This paper shows an overview of a virtual environment implementation of open-source vehicular traffic simulation software. This work contributes to this problem with a proposal of a pipeline that uses a set of open-source tools and integrates them in order to get a fully working traffic simulation in a virtual environment.

The rest of this paper is organized as follows. Section 2 presents selected previous works that had contributed to deal with the problem that motivated this work. Section 3 explains the proposed pipeline, describing in detail the whole pipeline that must be followed to achieve the expected result. Section 4 shows our study case that followed the pipeline described in Section 3. Finally, Section 5 contains the finals remarks, open problems, and perspectives of this work.

2. RELATED WORK

Studies about vehicle mobility, and their validation using computational tools, have background from some decades ago, as first example, a model for simulating vehicular traffic on multi-lane arterial roads named MULTISIM [4] presented in the 86, and which refers to articles from previous studies of similar validation approaches.

In the 90s, big companies from the automotive industry started to show inter-est on implement simulation systems, similar to already existing aircraft flying and submarine simulators. Consequently some simulators with these purpose were created by Volkswagen, the VTI (Swedish Road / Transport Research Institute) and Daimler-Benz [8].

Since the limitations imposed by the market prices and the technological con-straints these developments could not be commercially accessible, as they were only used to obtain data to develop new products and improve, some time later started to appear new proposals based on simpler computational interfaces, as for example the SIRCA simulator, developed on the Valencia University [1]. Also the SIMUVEG driving simulator and the truck training simulator PREVISIM-SICAM, both also from the Valecia University [11].

On the last years the most popular focus on these type of technologies have been the training tools for autonomous driving vehicles, due to the enormous range of situations able to reproduce on a virtual environment [6,7,9]. Semantic 3D models [12],for example, can already be produced by open-sourced data.

Regarding open-sourced vehicular simulation packages, there are three main options: *SUMO*, *CARLA*, and *AIRSIM*. SUMO, while useful, is a 2D vehicle simulation package that does not suit a 3D environment well. On the other hand, AIRSIM [10] and CARLA [2] are 3D driving and traffic simulation tools, focused on autonomous driving training, but with an enormous variety in functionality which let them be used for different projects.

These two were implemented on videogame engines as Unity 3D and Unreal Engine, reason why they can be used easily to create immersion experiences on virtual environments.

These tools are powerful but one lacks the features that the other has, and as such, a merging of the two will be necessary and further explained in the coming sections.

3. PROPOSED PIPELINE

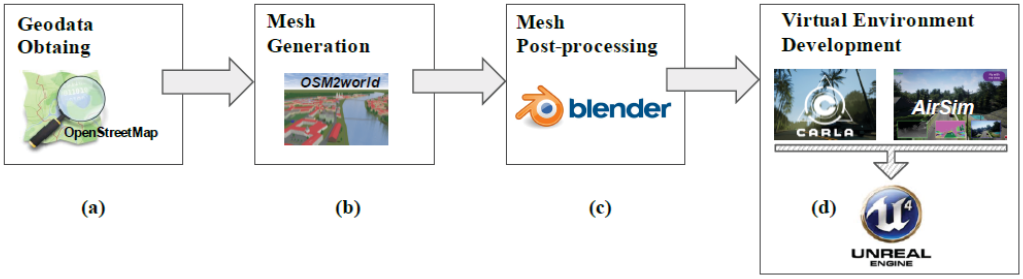


Fig. 1. Proposed Pipelines for the implementation of a simulation of vehicular traffic with open source software.

This work shows our version of a pipeline for the implementation of a simula-tion of vehicular traffic with open-source software. Figure 1 Depicts in a flowchart the sequence of steps that make up the proposed workflow. The first step (a), deal with the obtention of the necessary geodata, then (b) creates geometry from said data. The next step (c) makes the necessary interventions in the created ge-ometry, and finally, (d) explains the implementation in the virtual environment. The following subsections describe in detail each step of the proposed pipeline.

3.1 GEODATA OBTAINING

In order to create a virtual environment based on real-world data, the obtention of geodata is necessary and is the starting point of the process. It must be decided what area wants to be represented and then choose the software to get said data. In open-source tools this data is generally obtained in an XML file format, and usually has the following information: amount of roads, size of roads, the direction of roads, position of roads and amount of buildings in the selected area.
This information provides a starting point to generate a virtual representation of the city target.

3.2 MESH GENERATION

With the geomap data, it is necessary now to build geometry from said infor-mation. This geometry will populate the virtual environment and the intent is to result with a similar environment as the physical version.

Different techniques can be used to achieve this. Based on the format of the data that has been acquired, open-source tools can be used to automate the process. However, if no tool is present to interpret any particular geodata, then the mesh generation can also be done by hand by trained $3D$ artists.

3.3 MESH POST-PROCESSING

The obtained mesh from the previous step must be intervened in order to achieve two objectives. First, the geometry must be optimized to be able to run in a virtual environment, and second, the mesh has to be textured and rendered accurately in order to simulate the chosen city environment effectively.

In order to optimize the geometry, every mesh must have low polygon count, clean geometry, non-overlapping faces, correct normal orientation per face, and a simple UV unwrap for correct texturing. Once this

is done, all the geometry must be textured. This includes roads, buildings, side-walks, and everything that will be used in the environment. Open-sourced images can be used for this purpose. However, if the intent is to be as similar as possible to the real environment, it is recommended to find very similar textures to the ones that are found in the real environment.

3.4 VIRTUAL ENVIRONMENT DEVELOPMENT

Once the mesh is cleaned, the virtual environment can be set up. In this step, it is necessary to note that there are already a couple of open-souce tools that assist vehicular simulation for virtual environments, as stated before. However, whichever of these tools are chosen, a game engine in needed for them to run, so one must be picked. Popular choices are Unreal engine or Unity. Now that the engine is chosen, the open-source tool can be loaded in the engine and ready for intervention. Finally the generated geometry must be placed in the virtual world, and with this, the entire environment is set up.

4 STUDY CASE

Based on the proposed workflow, our study case utilized some particular tools on each step that will be expanded upon in the following subsections.

4.1 GEODATA OBTAINING: *OPENSTREETMAP*

Due to the desire that the users of this simulation would interact in a familiar environment, real geographic data must be used to the creation of a virtual map.

OpenStreetMap [5] is an open source project whose objective is to create a set of geographic data of free usage and completely editable. It has a community that makes GeoData accessible in a variety of format types for different uses. In Fig. 2 we can observe the chosen area to create the map for this simulation, a small sector of the north of Bogotá city. Exporting this data from *OpenStreetMap* generates a platform-specific *XML* file that can be worked on in the next step.

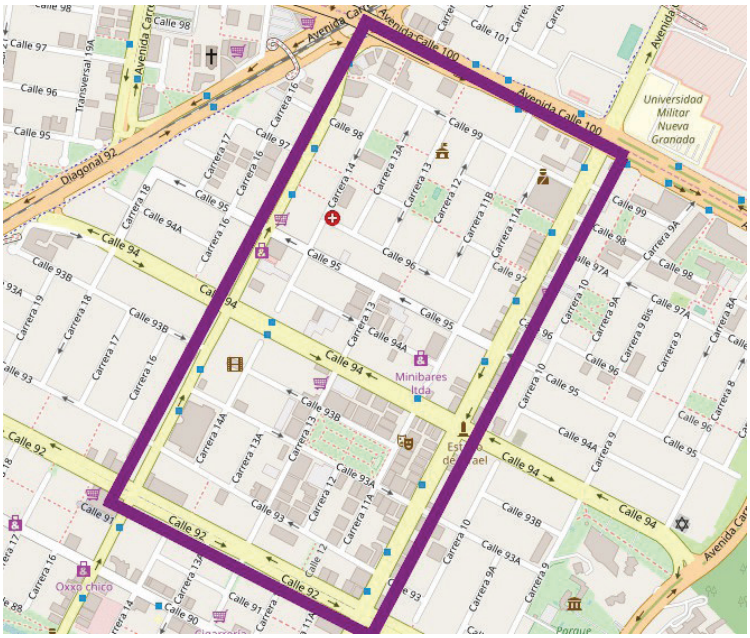


Fig. 2. Geographic space for the creation of the virtual environment map. Located on Bogota, Colombia, between the 97th and 100th street, 11th and 15th diagonals.

4.2 MESH GENERATION: *OSM2WORLD*

OSM2World (Open Street Map to word) is a java based, open-sourced solution that can take as input the OSM generated XML and output a $3D$ mesh that is built automatically from the data. The mesh includes roads, buildings and side-walks. Fig. 3 shows initial result of the implementation of the basic *OSM2World* map, which shows the same exact layout as the area that was selected from *OpenStreetMap* (Fig. 2).

4.3 MESH POST-PROCESSING: *BLENDER 3D*

The auto-generated mesh from the previous step had some problems in its geom-etry. *Blender 3D* is an open-source $3D$ package that allowed us to manipulate and correct the mesh, and it was used successfully in similar projects [3,12]. Once that was done, every mesh was textured utilizing UV unwrapping techniques with similar colors and textures to the real environment. The applied textures had a $1080p$ resolution, in order to aid with the performance of the simulation. Buildings, streets and stop lights were made to resemble the real environment as much as possible, Fig. 4 shows this result.

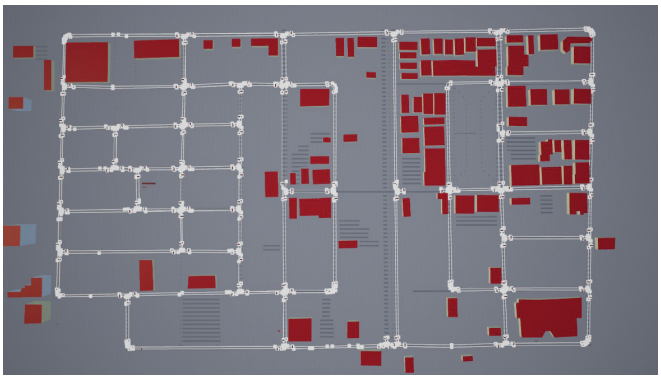


Fig. 3. Bogota streets implemented inside the game engine. The routes for the au-tonomous vehicles (white) covers all streets where a vehicle can move.

4.4 VIRTUAL ENVIRONMENT DEVELOPMENT: *CARLA + AIRSIM*

As previously stated, there are different kinds of software for traffic simulation, since for our case of study it must be open-source, the number of options to work with gets reduced. *SUMO*, *CARLA* and *AIRSIM* are the more suitable and the most known options as open software, but only *CARLA* and *AIRSIM* were selected because of the need to integrate these tools in a three dimensional virtual environment, since SUMO is a $2D$ tool.

AIRSIM [10] and *CARLA* [2] are driving and traffic simulation tools, focused on autonomous driving training, but with an enormous variety in functionality which let them be used for different projects. In reality, while both are publicized to have the same functions, in practice one excels over the other in different areas. *AIRSIM* has a robust physics engine for driving a car, but a lack-luster traffic simulation engine [10]. On the other hand, *CARLA*'s driving physics are not as polished as their vehicle simulation engine.

Fig. 4. The final mesh implemented in the virtual environment. It was designed to be as faithful to the real place as possible.

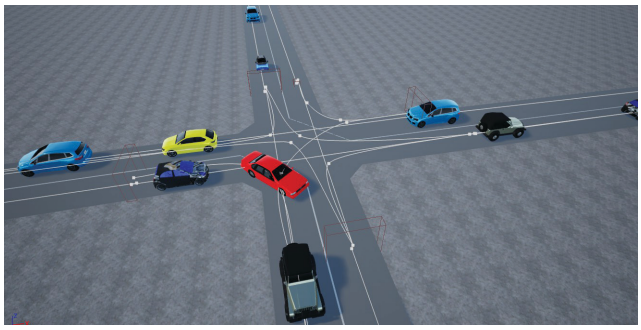


Fig. 5. Autonomous vehicles moving across a city intersection. They keep distances between each other and slows down as required.

Understanding this fact, it was necessary to combine both these tools in order to have one single cohesive simulation [2].

When choosing the game engine, since both of these tools were built natively on top of the Unreal engine, this was the engine of choice for the simulation in order to mix the two tools together. It was then convenient to implement the *CARLA* traffic simulation with this simple map and layout, Fig. 5 shows the result of this implementation. A spline based system was set up in every road for the autonomous vehicles to follow.

Then, the geometrically-correct map was also implemented in unreal engine. Also there are added traffic lights, signals and speed limits on the different places where they exist on the real physical space, with the purpose of increase the immersion through the familiarity of the user and the real place. The vehicles are able to interact with all of these signals and limits.

Also there is added traffic lights, signals and speed limits on the different places where they exist on the real physical space, with the purpose of increase the immersion through the familiarity of the user and the real place. The vehicles are able to interact with all of these signals and limits.

With these systems in place, the last necessary item was to include a user-driveable vehicle. *AIRSIM* was implemented alongside *CARLA* for this purpose. This tool has a great car physics system and is compatible with specialized vehicle simulation hardware, as such as car steering wheels or external controls,

to increase the user immersion level, allowing the user to drive as normally would. Fig. 6 depicts an user utilizing all these implemented systems in unison to interact with the virtual environment.

5. CONCLUSIONS AND FUTURE WORK

In this work we presented a pipeline divided in four steps that allows us to obtain geodata, generate a mesh from the data, process the mesh and implement these geometries with current vehicle simulation packages. All of these tools are open-source.

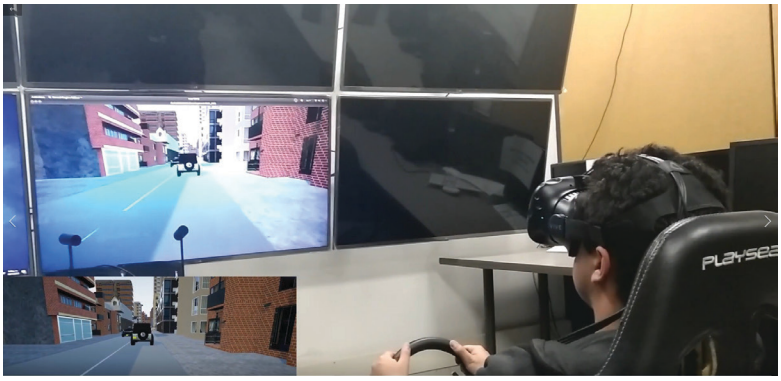


Fig. 6. User driving the vehicle utilizing a physical wheel. The steering and pedals work as expected and the familiar routes quickly inmerse the user in the world.

Although the *CARLA* tool is focused on the virtual environments simulation for autonomous vehicles artificial intelligence training, the fact that it is open source opens the possibility of it being used without commercial restrictions. The simulator systems are well implemented, giving tools to create different scenarios within successful simulations with differents purposes, as for our study case, immersion and interaction experiments.

Also all these system flexibility allowed to implement the desired virtual sce-nario, even if the development had to be done without certain external tools that the *CARLA* team had implement to that end, but where behind a commercial licence from a third party. *AIRSIM* systems will be used for the user vehicle manipulation, using hardware as steering wheels and pedals to attain immersion and interaction for the user.

As this simulation environment implemented with *CARLA* and *AIRSIM* will let us develop more experiments without having to implement artificial intelligent traffic simulation systems, even if there are some limiting things, these tools are of enormous help to future projects.

To increase immersion, new building geometry, streets and paths will be created, expanding the simple environment generated using *OpenStreetMap*, re-specting the orientation, quantity and size of the paths already implemented.

Finally, new visualization tools will be considered, as VR equipment, or a screen high resolution matrix, that will allow an even bigger immersion to the user in the virtual world.

REFERENCES

1. Bayarri, S., Fernandez, M., Perez, M.: Virtual reality for driving simulation. Communications of the ACM **39**(5), 72–76 (may 1996). https://doi.org/10.1145/229459.229468
2. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An Open Urban Driving Simulator. In: Proceedings of the 1st Annual Conference on Robot Learning. pp. 1–16 (oct 2017), http://proceedings.mlr.press/v78/ dosovitskiy17a.html
3. Dounas, T., Sigalas, A.: Blender, an Open Source Design Tool: Advances and In-tegration in the Architectural Production Pipeline. In: The New Realm of Archi-tectural Design [27th eCAADe Conference Proceedings. pp. 737–744 (2009), http: //papers.cumincad.org/cgi-bin/works/U=P=http/Show?ecaade2009{_}025
4. Gipps, P.: Multsim: a model for simulating vehicular traffic on multi-lane arte-rial roads. Mathematics and Computers in Simulation **28**(4), 291–295 (aug 1986). https://doi.org/10.1016/0378-4754(86)90050-9
5. Haklay, M., Weber, P.: OpenStreetMap: User-Generated Street Maps. IEEE Per-vasive Computing **7**(4), 12–18 (oct 2008). https://doi.org/10.1109/MPRV.2008.80
6. Mao, T., Wang, H., Deng, Z., Wang, Z.: An efficient lane model for complex traffic simulation. Computer Animation and Virtual Worlds **26**(3-4), 397–403 (may 2015). https://doi.org/10.1002/cav.1642
7. Miha˘i¸ta˘, A.S., Dupont, L., Camargo, M.: Multi-objective traffic sig-nal optimization using 3D mesoscopic simulation and evolutionary algo-rithms. Simulation Modelling Practice and Theory **86**, 120–138 (aug 2018). https://doi.org/10.1016/J.SIMPAT.2018.05.005
8. Nordmark, S.: Driving simulators, trends and experiences. In: RTS'94" Driving Simulation" Conference (1994), https://trid.trb.org/view/550128
9. Qiu, W., Shangguan, W., Cai, B., Chai, L.: Research on Traffic Simulation Sce-nario Construction and Fidelity Evaluation Method under Environment of i-VICS. In: 2019 Chinese Control Conference (CCC). pp. 7155–7160. IEEE (jul 2019). https://doi.org/10.23919/ChiCC.2019.8866169
10. Shah, S., Dey, D., Lovett, C., Kapoor, A.: AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In: Field and Service Robotics. Proceedings in Advanced Robotics, pp. 621–635. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-67361-5 40
11. Valero-Mora, P., Pareja, I., Pons, D., Sa´nchez, M., Montes, S.A., Ledesma, R.D.: Mindfulness, inattention and performance in a driving simulator. IET Intelli-gent Transport Systems **9**(7), 690–693 (sep 2015). https://doi.org/10.1049/iet-its.2014.0172
12. Wendel, J., Simons, A., Nichersu, A., Murshed, S.M.: Rapid development of seman-tic 3D city models for urban energy analysis based on free and open data sources and software. In: Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics - UrbanGIS'17. pp. 1–7. ACM Press, New York, New York, USA (2017). https://doi.org/10.1145/3152178.3152193

SISTEMA COLABORATIVO DE MEDICIÓN DE PARÁMETROS AMBIENTALES BASADO EN IOT

MAUREEN PARRA 1, EDWARD GUILLEN 2, FRÉDÉRIC LE MOUËL 3 Y OSCAR CARRILLO 4

1-2 UNIVERSIDAD MILITAR NUEVA GRANADA, BOGOTÁ, COLOMBIA

3-4 INSTITUT NATIONAL DES SCIENCES APPLIQUÉES (INSA), LYON, FRANCIA U1401167@UNIMILITAR.EDU.CO EDWARD.GUILLEN@UNIMILITAR.EDU.CO

FREDERIC.LE-MOUEL@INSA-LYON.FR OSCAR.CARRILLO@CPE.FR

RESUMEN.

Las alertas ambientales declaradas en Bogotá, los informes mundiales sobre la situación actual del medio ambiente y los efectos de la contaminación sobre la salud de las personas son las principales causas por las que el grupo de investigación GISSIC (Grupo de Investigación en Seguridad y Sistemas de Comunicación) decidiera elaborar un proyecto que pudiera entregar una solución tecnológica alternativa, en tiempo real y a una escala importante.

El despliegue de nodos IoT capaces de realizar las medidas de las variables con-taminantes que se encuentran suspendidas en el aire se realizará inicialmente en el campus Nueva Granada que se encuentra ubicado en Cajicá y posteriormente a las demás sedes de la universidad Militar Nueva Granada. Los datos recolecta-dos serán enviados a una interfaz de usuario en la que se podrá visualizar el com-portamiento de las variables seleccionadas. Estos datos serán públicos para que la comunidad tenga acceso a esta información y pintos de voluntariado podrán acrecentar la red para hacerla colaborativa.
La arquitectura que se está desarrollando se compone por sensores distribuidos y una plataforma escalable con capacidades de administración y gestión remota, así como procesamiento en edge computing. La plataforma tendrá la capacidad de integrarse con la plataforma de administración de servicios que se está imple-mentando en el WIRID LAB y las plataformas de Future Internet of Things de INSA en Francia.

Palabras claves: Calidad del aire, Datos abiertos, Escalable, Autoconfigurable, Bajo costo.

1 INTRODUCCIÓN

Hoy en día estamos viviendo el posicionamiento de nuevas tecnologías en donde se integran dispositivos como sensores inalámbricos y servicios como Computación en la nube. La tecnología es ahora más interactiva con los usuarios finales, entregando solu-ciones de conectividad que facilitan la vida de estos. Tecnologías como el Internet de las cosas están revolucionando el mundo ya que son capaces de ofrecer modelos tecno-lógicos en el área del transporte, salud, hogar, entre otras. La contaminación es uno de los principales problemas que afecta la salud de las personas ya que el medio ambiente y la salud humana están directamente relacionadas. El material particulado más dañino es de un tamaño menor a 10 micros, este es capaz de producir problemas cardiovascu-lares y respiratorios. Por otra parte, los gases de efecto

invernadero, principalmente el Dióxido de carbono –CO2, es el responsable del cambio climático, por estas razones se quiere desarrollar un proyecto orientado al internet de las cosas que entregue diferentes datos que informen sobre el estado de la calidad del aire. El modelo de la plataforma propuesta en el desarrollo del presente artículo permitirá la implementación y conexión de nodos IoT medidores de la contaminación creados por investigadores del grupo GISSIC y adquiridos por usuarios externos que quieran conectarse a esta plataforma de servicios y compartan los datos recolectados mediante sus nodos IoT a la comunidad. Este proyecto está basado en la plataforma experimental FIT/IoT-LAB que ofrece ser-vicios y aplicaciones IoT a gran escala mediante el laboratorio IoT-LAB. Los principa-les servicios que soporta es una arquitectura heterogénea, es decir, hardware, topologías de red, protocolos y bibliotecas orientadas a IoT. La gestión y administración de los experimentos, además, la realización de estas actividades mediante herramientas de vi-sualización y ejecución. Este laboratorio fue implementado con la intención de avanzar en pruebas del sector de IoT, por esta razón es de acceso libre para la academia, para empresas y para enseñanza [1].

2 ANTECEDENTES Y ESTADO DEL ARTE

Existen diferentes proyectos en el mundo sobre la medición a gran escala de la calidad del aire mediante trabajos orientados a internet de las cosas, a continuación, se explica-rán las principales características de cada uno.

En Colombia, existe un sistema de monitoreo del aire implementado en la ciudad de Medellín (Antioquia), este sistema es llamado SIATA (Sistema de Alerta Temprana de Medellín y el Valle de Aburrá) es un conjunto de red de monitoreo de variables am-bientales creado por un grupo de investigación en ciencia y tecnología. Una de las redes que se desarrollaron es responsable de monitorear la calidad del aire en el Valle de Aburrá. Las principales variables medidas por estas estaciones son: ozono, óxidos de nitrógeno, monóxido de carbono, PM 10 y PM 2.5.

Las 43 estaciones instaladas en la ciudad recolectan los datos tomados por los sen-sores y los envían a los servidores que están en la planta, allí procesan y analizan esta información para establecer si son datos correctos y confiables. Según los resultados del análisis de datos, el equipo tomará una decisión, por ejemplo, si existe una situación que puede afectar la salud pública, se seguirá un protocolo establecido [2].

AIRBEAM: SHARE & IMPROVE YOUR AIR (AIRBEAM: COMPARTE Y MEJORA TU AIRE) es un proyecto desarrollado por la empresa HabitatMap, una com-pañía ambiental sin ánimo de lucro que se encuentra ubicada en Nueva York. Habitat-Map lanzo un dispositivo llamado AIRBEAM que realiza medidas sobre la calidad del aire del entorno, el dispositivo cuenta con sensores que son capaces de detectar partí-culas o material articulado que se encuentra suspendido en el aire como lo es el PM 2.5, concentraciones de gases como Dióxido de Nitrógeno y Monóxido de Carbono, ade-más, mide variables ambientales como la temperatura y la humedad.

AIRBEAM está conectado con una plataforma llamada AirCasting en la que se re-colectan datos de cada AIRBEAM y se publican para que la comunidad los conozca y así se pueda evitar o mejorar daños en la salud.

La forma de funcionamiento de este dispositivo es mediante una luz led que produce la dispersión de las partículas que se encuentran en el aire, esta dispersión es registrada por el detector y con esto se estima cierta concentración o cantidad de partículas en el aire. El dispositivo medidor envía las medidas tomadas mediante Bluetooth a la aplica-ción AirCasting que estaría ubicada en el Smartphone del usuario y al final de cada sesión se envían el total de datos recolectadas a la página web de la plataforma para formar el mapa que indica las zonas con mayor concentración de PM 2.5.

Después de la recolección de datos, el sensor recibe respuesta de la aplicación y el dispositivo se ilumina según la concentración de la medida indicando con el verde, amarillo, naranja o rojo para alta concentración [3].

IoT ha permitido la creación e integración de diferentes dispositivos tecnológicos para la prestación de servicios en la sociedad. La variedad de aplicaciones de esta tec-nología hace que incremente el número de usuarios en esta red heterogénea. Estos nue-vos dispositivos creados deben ser interconectados, aunque su capa física y su protocolo de comunicación varíe. EL-MOUGY y los demás autores explican cómo podría for-marse una arquitectura que integraría dispositivos IoT conformados por diferentes tipos de sensores y como esta arquitectura podría ser escalable y personalizada, según sus aplicaciones, para cada usuario [4].

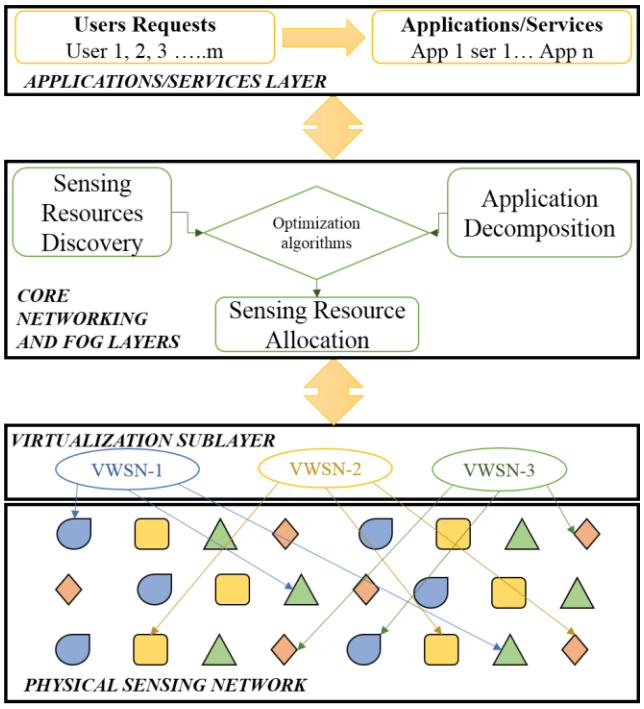


Fig. 1. Arquitectura IoT con hardware heterogéneo [4].

El aumento de los usuarios en las tecnologías como IoT genera la creación de arqui-tecturas como la observada en la figura 1, en donde se busca implementar una virtuali-zación de recursos que permite reducir costos, facilitar el escalamiento y reusar hard-ware ya instalado. A demás se crea una capa de aplicación que se configura con la capacidad de responder mediante diferentes a los servicios que solicitan los usuarios. La capa física y la capa de aplicación está interconectada mediante la capa Fog permite la escalabilidad y soporte en tiempo real para aplicaciones de IoT y análisis en paralelo de los datos recolectados [4].

Los servicios desarrollados para IoT actualmente utilizan middleware para el esta-blecimiento de comunicación entre la infraestructura en la que se encuentran. Cada ele-mento IoT trae por defecto su middleware configurado y una API que el proveedor entrega para poder programar el dispositivo, esto hace difícil tener acceso a varios re-cursos de IoT porque no pueden conectarse a middleware diferentes. En la plataforma propuesta que se puede observar en la figura 2, se muestra un método de acceso unifi-cado que puede superar los problemas que presenta la heterogeneidad de los dispositi-vos IoT [5].

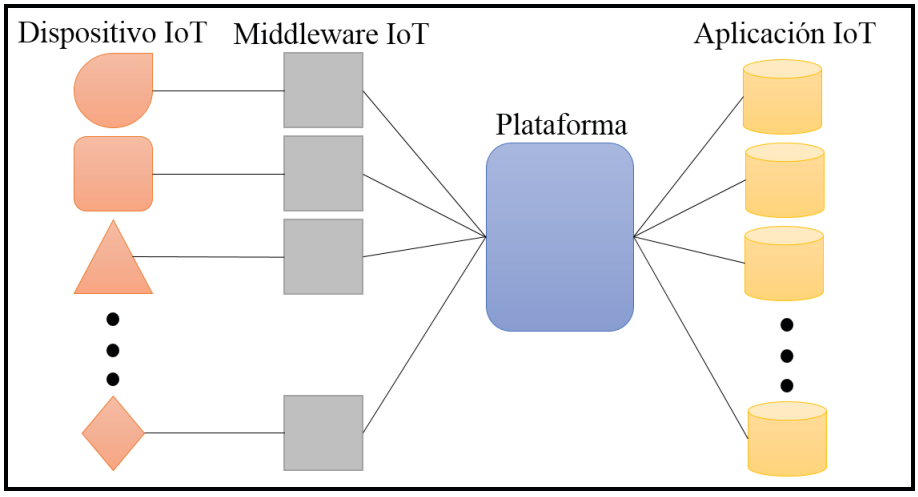


Fig. 2. Arquitectura de plataforma de servicios con hardware heterogeneo [10].

En la plataforma abierta de servicios desarrollada cada aplicación IoT no necesita conocer el middleware de cada dispositivo heterogéneo ni a que recursos debe acceder. En esta arquitectura propuesta las aplicaciones IoT solicitan apertura a la plataforma de servicios y el procedimiento restante lo realiza la plataforma. Esta plataforma actúa como un intermediario para realizar la solicitud específica a cada middleware IoT dife-rente [5].

- Uso completo y fácil acceso al recurso IoT y a los datos recolectados.
- Conexión y comunicación fácil con los recursos IoT.
- Optimización y distribución de las aplicaciones.

FIT IOT-LAB

La plataforma FIT (Future Internet of the Things) es una herramienta que comple-menta el IoT-LAB que ayuda a la realización de experimentos mediante conexión re-mota con los nodos ya instalados. Los nodos implementados permiten realizar experi-mentos libres para cada usuario y medir variables como el consumo de energía, la in-terferencia de señal y métricas de evaluación de la red como, por ejemplo, rendimiento, delay, entre otras.

Los resultados de los experimentos ejecutados con los nodos pueden ser analizados y evaluados [6].

El FIT IoT-LAB posee 2728 nodos inalámbricos y 117 robots móviles para la eje-cución de los experimentos en IoT, estos elementos están distribuidos en 6 diferentes lugares en Francia y esta infraestructura lo hace el laboratorio remoto de pruebas y experimentos más grande del mundo.

Los nodos IoT están ubicados en diferentes topo-logías de red y poseen una variedad de sensores inalámbricos que son completamente programables.

Existe un alto nivel de accesibilidad a las puertas de enlace de cada nodo, lo que permite controlar, monitorear y programar los nodos, además, también se puede movilizar los robots y ver como se afectan los protocolos de comunicación con estos cambios en tiempo real [1].

Los nodos IoT están conectados físicamente mediante un backbone que les provee la energía de alimentación y los conecta a los servidores [1]. Existen 3 principales com-ponentes de los nodos en el IoT-LAB:

1. ON (Open Node): es el dispositivo de bajo consumo que el usuario reprograma, ade-más está conectado directamente a la puerta de enlace mediante un puerto serial [1].
2. GW (Gateway): o puerta de enlace está conectado al backbone mediante una cone-xión ethernet, además de esto, monitorea y reporta la actividad realizada por el Open Node (ON) [1].
3. CN (Control Node): o nodo de control se encarga de coordinar la reprogramación que realiza el Open Node, puede apagar, encender y reiniciar el nodo, así mismo, elegir la fuente de alimentación, entro otras actividades [1].

Existen diferentes tipos de nodos implementados en el laboratorio, como los siguientes: WSN430 ON (862MHz), WSN430 ON (2.4GHz), M3 ON, A8 ON y Generic host node (no ON) y dos tipos de robots como: Turtlebot y Wifibot [1].

WIRID LAB

El WIRID LAB es un laboratorio en el que se realizan proyectos de comunicación inalámbrica. Está conformado por diferentes elementos que permiten una comunica-ción inalámbrica y su respectivo análisis.

Algunos equipos son los de radio definido por software como, por ejemplo, las USRP, también se encuentran en el laboratorio los analizadores de espectro, diferentes tipos de antenas que son clasificadas por alta o baja frecuencia de operación y generadores RF que también actúan en diferentes frecuencias [7].

Estos equipos están distribuidos a lo largo del laboratorio y permiten simular esta-ciones de transmisión y recepción como estaciones base de comunicaciones móviles. En la infraestructura del WIRID LAB se instalaron varios puntos de conexión a internet por ethernet para lograr comunicar los elementos que no se pueden conectar por WIFI.

La conectividad que hay en el laboratorio permite el despliegue de más elementos de comunicación y medición en una misma red que es administrada por una plataforma de gestión que ya se está implementando [7].

CALIDAD DEL AIRE

La contaminación en el aire en uno de los principales factores que afectan la salud de las personas y ha producido entre 6 y 7 millones de muertes prematuras, según el último informe de la ONU [8]. El material particulado que se encuentra en el aire puede ingresar a los pulmones y generar enfermedades como disminución en la función pul monar, enfermedades cardiovasculares, inflamación del pulmón, problemas respirato-rios y muertes por enfermedades cardiopulmonares.

Estas partículas se deben clasificar según su diámetro para poder ser monitoreadas. Existen tres principales categorías de PM, las PM 10 y PM 2.5 son partículas de diámetro grueso porque se encuentran entre 2.5 y 10 micrómetros, son inhalables y respirables, pero no se depositan en los pulmo-nes. Las PM 1.0 son partículas ultrafinas de un diámetro menor a 1 micrómetro [9].

Los gases son partículas de sustancias químicas que afectan la calidad del aire. La emisión de gases como por ejemplo el dióxido de carbono (CO_2) generan el efecto invernadero, la principal causa del

cambio climático. Esta emisión de gases es generada principalmente por la quema de combustibles para ser utilizados en la industria, la elec-tricidad y el transporte [8]. El sector del transporte es el principal contaminante de la atmosfera. Las partes más habitadas de las ciudades son muy afectadas por el tráfico que se encuentra siempre cerca del entorno de las personas afectando su estado de salud. Como consecuencia de esto, se producen enfermedades como asma, alergia, bronquitis, entre otros [10].

La medición de estas variables contaminantes permite la creación de bases de datos que ayudan a la toma de decisiones en diferentes aspectos que generan la contaminación en el medio ambiente.

IoT es una tecnología que permite la conexión de diferentes dis-positivos inteligentes y el intercambio de datos por medio de la red. El internet de las cosas se ha utilizado en diferentes sectores como la vivienda, el transporte, sistemas de seguridad, entre otros. El monitoreo de la calidad del aire se ha convertido en una acti-vidad muy importante en diferentes ciudades debido al efecto negativo de la contami-nación en la salud de la comunidad [11].

3 PROPUESTA DE ARQUITECTURA

La herramienta que se está utilizando para administrar los dispositivos es Kubernetes, una plataforma que se utiliza para gestionar servicios y recursos basados en contenedores, además, facilita la configuración y automatización de un sistema implementados en cluster (conjunto de nodos).

Al ser una aplicación en la nube se puede implementar en sistemas escalables y de rápido crecimiento. Esta plataforma presta diferentes servicios entre ellos: la administración de contenedores, además de esto, es portable para servicios de computación en la nube y otros microservicios que pueda ejecutar.

En la arquitectura de Kubernetes se encuetran dos principales objetos el master y el nodo, el master se encarga de controlar y administrar los nodos, además tiene un conjunto de procesos que solo el puede ejecutar y dentro de cada nodo se encuentran los pods, que son conjuntos de contenedores que contienen diferentes aplicaciones [12].

Existen dos tipos de pods, los pods de servicio y los pods de trabajo. Los pods de servicio o service pods se están ejecutando todo el tiempo en el segundo plano del cluster, a este pod se asociacian dos métricas la disponibilidad y lautilización del servicio. Los pods de trabajo se encargan de ejecutar o terminar una tarea, a este pod se asocian las métricas de desarrollo y tiempo de ejecución. El de los contenedores que se encuentren dentro del pod determinarán el tiempo de vida del mismo [13].

Los contenedores y las máquinas virtuales se diferencian principalmente por los re-cursos físicos que cada herramienta utiliza en el equipo. Las máquinas virtuales realizan una copia completa de su sistema operativo y de los recursos de hardware que necesitan para ser ejecutadas.

Por otro lado, el contenedor se ejecuta en la capa de aplicación, es decir, las copias que realiza para su ejecución son de menor tamaño como, por ejemplo, el peso de la imagen de un contenedor se encuentra en decenas de MB (MegaBytes). Así mismo, la ejecución de varias aplicaciones al tiempo con una baja utilización de recursos hace que esta unidad de software sea más optima [14].

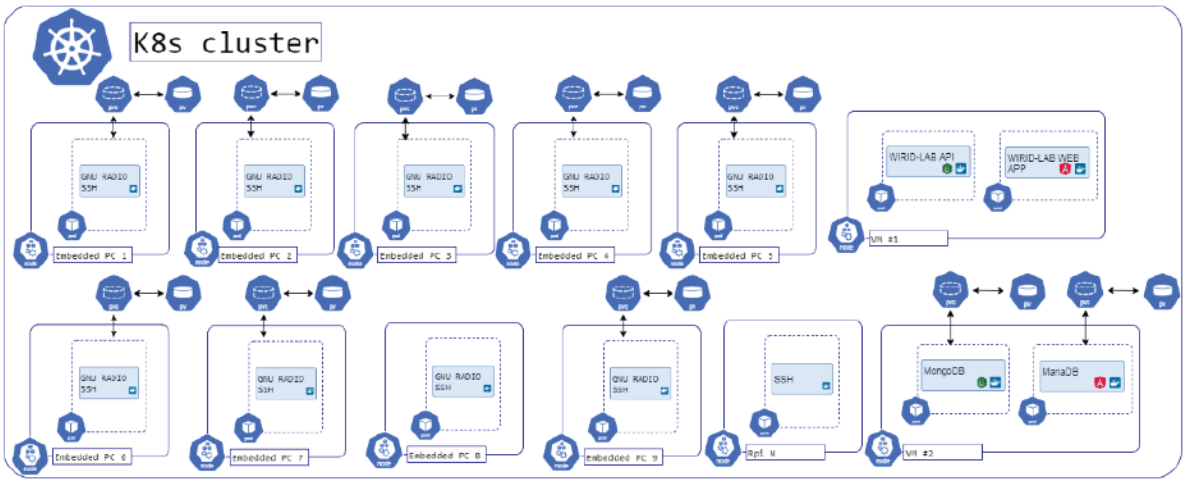


Fig. 3. Arquitectura en Kubernetes del WIRID LAB

En la arquitectura modelada del WIRID LAB en Kubernetes que se puede observar en la figura 3. Se encuentra un cluster con 10 nodos y cada nodo es un computador embebido en el cual se encuentran diferentes aplicaciones como SSH, la API del labo-ratorio, el servidor web o GNU para las USRPs, entre otros.

Todos los nodos dentro de la infraestructura se pueden comunicar entre sí y además de esto poseen un Volumen Persistente (PV) que les permite guardar información en el equipo en donde se está alojando esto se realiza mediante una petición de almacenamiento llamada Reclamo de Volumen Persistente (PVC) [12].

En las principales aplicaciones instaladas en los no-dos se encuentra MongoDB, una base de datos creada para aplicaciones modernas que provee soporte en cualquier escala y se caracteriza por tener alta disponibilidad, esca-labilidad, es decir, que tiene la capacidad de crecer según la arquitectura, ejecuta cargas de trabajo en el mismo clúster, puede distribuir los datos en dispositivos seleccionados y en cualquier zona geográfica, además es una base de datos portable que funciona correctamente en cualquier parte, además, esta puede ser implementada en la nube.

Es-tas características permiten que se pueda crear una plataforma que gestione datos ope-racionales respaldados por Mongo DB [15]. En la plataforma se utilizó para almacenar la API y la base de datos del WIRID LAB.

Maria DB es otra aplicación instalada en uno de los nodos creados y es utilizada para soportar la página web del WIRID LAB, esta aplicación es un servidor de bases de datos que convierte los datos estructurados en aplicaciones como, por ejemplo, un sitio web. Se caracteriza porque se puede desarrollar en entornos escalables y posee herra-mientas como el almacenamiento que lo hace útil en diferentes aplicaciones [16].

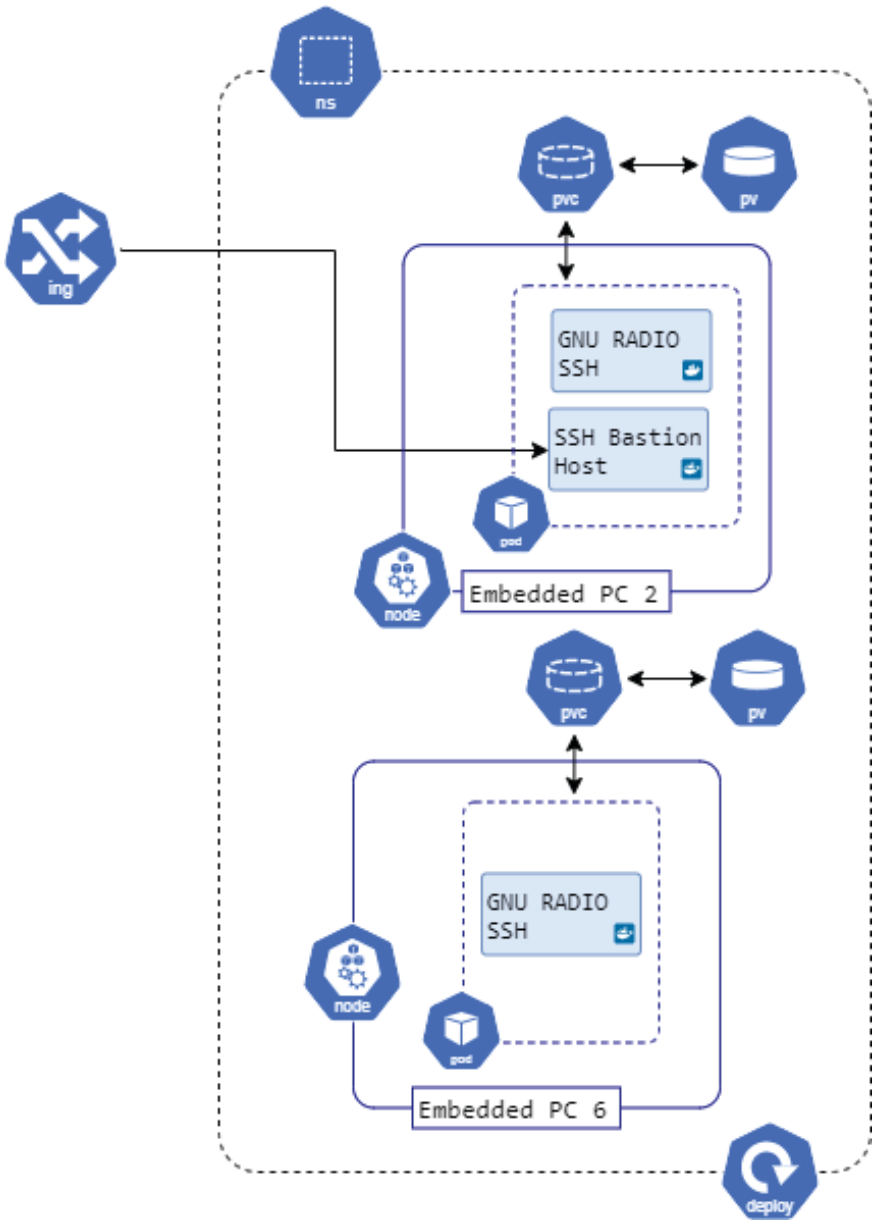


Fig. 4. Nodo de servicio de Kubernetes aislado.

En la figura 4, se pueden observar un nodo aislado de la arquitectura, una de las funcionalidades que también permite realizar Kubernetes, en este gráfico se encuentran nuevos elementos como Ingress que actúa como puente para exponer puertos que per-miten el uso de diferentes servicios que son externos al nodo. Name services o ns per-mite la separación de nodos por grupos y Deploy es una herramienta que se encarga de llevar al nodo a un estado deseado mediante una actualización [12].

Las características de este laboratorio permiten el despliegue de más equipos que sean capaces de enviar información a la plataforma de gestión, por esto, los investiga-dores del grupo GISSIC (Grupo de Investigación en Seguridad y Sistemas de Comuni-cación) crearon unos nodos IoT que realizan medidas sobre diferentes variables del entorno, estos dispositivos también están siendo agregados a la plataforma Kubernetes para poder ser administrados.

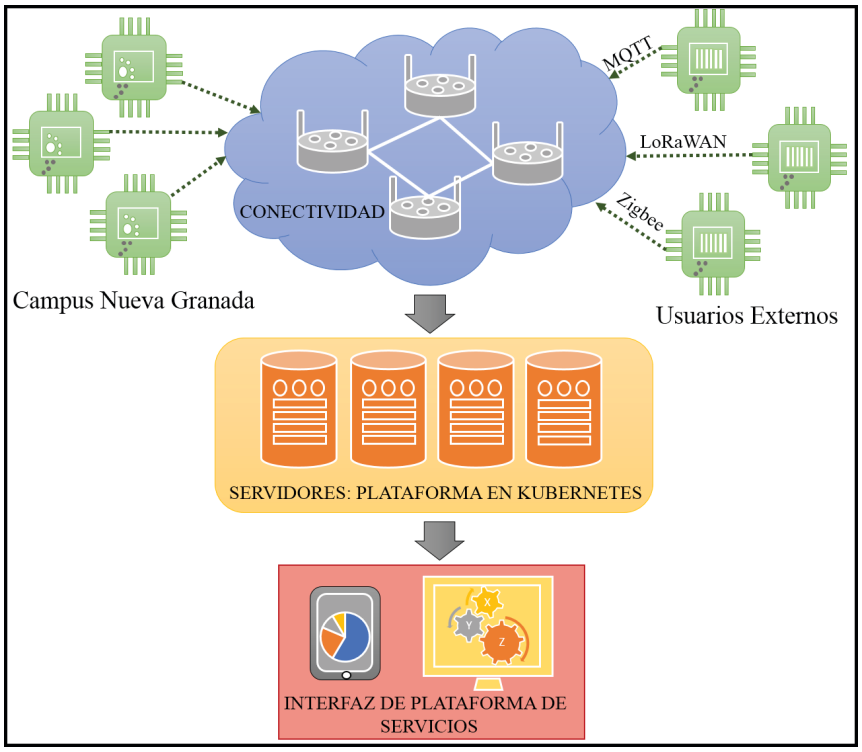


Fig. 5. Arquitectura propuesta de un sistema IoT de medición de calidad del aire con hardware heterogéneo

Este proyecto de despliegue de nodos IoT medidores de calidad del aire en el Cam-pus Nueva Granada nace de la necesidad de aportar a la sociedad una alternativa de datos sobre contaminación en el aire diferente a las estaciones que implementó el estado porque el proyecto está orientado a que la comunidad tenga accesibilidad a la plata-forma. Aunque la plataforma esté implementada en esta sede de la universidad estará abierta para que cualquier ciudadano conecte su nodo medidor de contaminación y aporte datos a la plataforma.

En la figura 5, se puede observar una imagen de la arquitectura propuesta en donde se puede encontrar el hardware o los sensores utilizados para la medición de las varia-bles ambientales contaminantes como PM 10, PM 2.5, gases específicos, entre otros.

Como ya se ha mencionado, los medidores adquiridos por los usuarios tendrán la capa-cidad de conectarse a la plataforma ya que estos se pueden configurar para que esta-blezca una comunicación mediante cualquier protocolo como los que se explican más adelante. Una vez conectados al servidor, los usuarios tendrán acceso a la plataforma de servicios en donde pueden publicar los datos tomados con los medidores IoT, ade-más pueden observar mediciones que otras personas han publicado y que los nodos instalados en la universidad también han realizado.

Posteriormente, se encuentra la in-terfaz gráfica de la plataforma de servicios que se refiere a la muestra de los datos re-colectados a la comunidad interesada en conocer el estado actual de la calidad del aire en su entorno.

La comunicación y el envío de datos del nodo a la plataforma se puede realizar mediante diferentes protocolos de comunicación como por ejemplo MQTT, LoRaWAN o Zig-bee, entre otros. A continuación, se explica el funcionamiento básico de cada uno de estos protocolos de comunicación.

1. MQTT (Message Queuing Telemetry Transport) o en español transporte de teleme-tría de colas de mensajes, es un protocolo que pertenece a la capa de transporte según el modelo OSI, así mismo este protocolo se encarga de transportar la información desde un extremo al otro. Este protocolo está diseñado para la transmisión de datos generados por dispositivos IoT mediante una red inalámbrica [17].

MQTT está basado en la interacción entre dos extremos llamados: suscriptor y publicador, en donde ninguno conoce la identidad del otro. Está conformado por dos partes como son: los clientes (suscritor y publicador) y Broker [17].

Los clientes MQTT son los dispositivos, nodos IoT, sensores o elementos con capacidad de conectarse a la red que se comunican o envían y reciben información que es transportada por este protocolo.

El Broker es la parte que se comporta como central de conexión de todos los clientes. Los datos intercambiados mediante el pro-tocolo MQTT son llamados mensajes los cuales están contenidos dentro de un Tó-pico, estos son los temas o divisiones de información general que se puede transmitir después de ser recolectada [17].

2.LoRaWAN es una tecnología que ofrece un servicio de comunicación en una larga cobertura y con un bajo consumo de energía en los dispositivos finales. LoRaWAN está basado en el CSS (chirp espectro ensanchado), que es una técnica en la que la frecuencia de la señal aumenta con el tiempo hasta que ocupa todo el ancho de banda de la señal. Básicamente LoRa es la tecnología base de la comunicación y LoRa-WAN se encarga de gestionar la comunicación establecida. Gracias a la modulación que realiza LoRa se pueden alcanzar grandes distancias de comunicación, esta tec-nología está estandarizada por la alianza LoRa y ha establecido aspectos como ad-ministración de dispositivos y mensajes, formato de las tramas generadas, acceso al medio, entre otros [18]. LoRaWAN provee funcionabilidades como la conexión de muchos nodos finales que se pueden comunicar por medio de uno o varios gateways de esta tecnología. Los gateways actúan como canales de transmisión de mensajes hacia servidores [19].

3. Zigbee es un protocolo de comunicación inalámbrico, se caracteriza por funcionar con un bajo consumo de energía a cortas distancias y con una tasa baja de datos, alcanzando los 250 Kbps en distancias de 10 a 70 metros. El consumo promedio de corriente es de 30 mA en modo standby [20]. Zigbee es un protocolo muy seguro en comparación a otras tecnologías como WIFI y se encuentra estandarizado en IEEE 802.15.4, además de esto, trabaja sobre las frecuencias 915 y 868 MHz con una lon-gitud de onda de 2.4 GHz, cuando se trabaja en la frecuencia más baja la potencia recibida es más alta pero la tasa de datos se ve limitada. La comunicación de este protocolo se basa en el esquema servidor y cliente, en la puerta de enlace se confi-gura el coordinador que cumple la función de servidor y otro subdispositivo de Zig-bee actúa como cliente, estos componentes se envían y responden mensajes de con-trol, consulta y estado [21].

El medidor de contaminación que se construirá en el grupo de investigación GISSIC para la formación de la base de datos de las medidas contaminantes del aire, debe con-formarse con sensores de bajo costo que midan los siguientes parámetros:

- Material Particulado: como se explicó anteriormente, este tipo de partículas afectan negativamente el estado de salud de las personas por esto es importante realizar me-didas de concentración de PM en el entorno en que se convive y con base a esto poder tomar y generar decisiones con respecto a los cambios que se deben hacer en la ciudad.
- Gas: una de las mediciones de gas que se debe realizar es la del Dióxido de carbono (CO_2), el cual es producido por los medios de transporte principalmente, el conoci-miento de la concentración de este gas puede ayudar a incentivar cambios en com-bustibles que se utilizan en el sector del transporte y a retirar los autos, buses, ca-miones de carga, entre otros, que por su estado de antigüedad generan

estos gases dañinos. Sin embargo, si el usuario quiere medir y compartir datos sobre otros gases como el metano, monóxido de carbono, entre otros puede incluir este tipo de senso-res en su medidor.

- Temperatura y humedad: Es importante conocer estos dos aspectos porque las me-didas varían según el cambio de la temperatura y humedad del ambiente y es nece-sario tener en cuenta estas características.
- Tarjeta de comunicación: una herramienta que le permita comunicar el medidor a la

plataforma y enviar los datos recolectados para la construcción y organización de la información que se mostrará a la comunidad.

4 CONCLUSIONES Y TRABAJO FUTURO

Finalmente, en este documento se realizó la propuesta de una plataforma de servicios de datos públicos relacionados con la calidad del aire en el entorno en donde se realicen las medidas, por esto, también se entrega la posibilidad de la conexión de usuarios ex-ternos a la plataforma implementada en el servidor del laboratorio y así lograr la for-mación de una amplia base de datos que pueda informar a las personas el estado am-biental actual mediante una página web.

Como trabajo futuro, se busca implementar la plataforma propuesta con base a la infraestructura del WIRID LAB que se creó en Ku-bernetes puesto que esta será la que soporte la conexión e integración de los nodos implementados en el Campus Nueva Granada y los adquiridos por los usuarios externos a la universidad.

REFERENCIAS

1. Adjih, C., Baccelli, E., Fleury, E., Harter, G., Mitton, N., Noel, T., ... & Watteyne, T. (2015, December). FIT IoT-LAB: A large scale open experimental IoT testbed. In 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), pp. 459-464, IEEE (2015)
2. SIATA Sistema de Alerta Temprana de Medellín y el Valle de Aburrá, https://siata.gov.co/sitio_web/index.php/calidad_aire, último acceso 2019/10/15
3. Kickstarter. AIRBEAM: share and improve your air, https://www.kickstarter.com/pro-jects/741031201/airbeam-share-and-improve-your-air, último acceso 2019/10/14
4. El-Mougy, A., Al-Shiab, I., & Ibnkahla, M. Scalable Personalized IoT Networks. Pro-ceedings of the IEEE, 107(4), pp. 695-710, IEEE (2019).
5. Park, D. H., Bang, H. C., Pyo, C. S., & Kang, S. J. Semantic open IoT service platform technology. In 2014 IEEE World Forum on Internet of Things (WF-IoT), pp. 85-88, IEEE (2014).
6. FIT FUTURE INTERNET OF TESTING FACILITY. FIT IOT-LAB: SENSING EMBEDDED MOBILE, https://www.iot-lab.info/, último acceso 2019/9/20
7. WIRID LAB, http://wirid-lab.umng.edu.co/, último acceso 2019/10/10
8. ONU: Medio Ambiente.: Programa de las Naciones Unidas para el Medio Ambiente (2019), Perspectivas del Medio Ambiente Mundial, GEO 6: Planeta sano, personas sa-nas, Nairobi. PNUMA (2019).
9. Marques, G., Roque Ferreira, C., & Pitarma, R. A system based on the Internet of Things for real-time particle monitoring in buildings. International journal of environmental re-search and public health, 15(4), 821 (2018).
10. Forehead, H., & Huynh, N. Review of modelling air pollution from traffic at street-level-The state of the science. Environmental Pollution, 241, pp. 775-786 (2018).
11. Kaivonen, S., & Ngai, E. Real-time air pollution monitoring with sensors on city bus. Digital Communications and Networks. (2019).
12. Kubernetes, https://kubernetes.io/es/docs/concepts/overview/, último acceso 2019/10/18
13. Medel, V., Tolosana-Calasanz, R., Bañares, J. Á., Arronategui, U., & Rana, O. F. Char-acterising resource management performance in Kubernetes. Computers & Electrical En-gineering, 68, pp. 286-297. (2018).
14. Docker: The Modern Platform for High-Velocity Innovation, https://www.docker.com/why-docker, último acceso 2019/10/18

15. MongoDB Architecture, https://www.mongodb.com/mongodb-architecture, último ac-ceso 2019/10/19
16. MariaDB Foundation: Supporting continuity and open collaboration, https://mari-adb.org/about/, ultimo acceso 2019/10/19
17. Kashyap, M., Sharma, V., & Gupta, N. Taking MQTT and NodeMcu to IOT: Commu-nication in Internet of Things. Procedia computer science, 132, pp. 1611-1618. (2018).
18. Van den Abeele, F., Haxhibeqiri, J., Moerman, I., & Hoebeke, J. Scalability analysis of large-scale LoRaWAN networks in ns-3. IEEE Internet of Things Journal, 4(6), pp. 2186-2198 (2017).
19. Barro, P. A., Zennaro, M., & Pietrosemoli, E. TLTN–The local things network: on the design of a LoRaWAN gateway with autonomous servers for disconnected communities. In 2019 Wireless Days (WD) (pp. 1-4). IEEE (2019).
20. Nugroho, E., & Sahroni, A. ZigBee and wifi network interface on Wireless Sensor Net-works. In 2014 Makassar International Conference on Electrical Engineering and Infor-matics (MICEEI) pp. 54-58. IEEE (2014).
21. Pan, G., He, J., Wu, Q., Fang, R., Cao, J., & Liao, D. Automatic stabilization of Zigbee network. In 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD) pp. 224-227. IEEE (2018).

A PERCEPTUAL CALIBRATION METHOD TO AMELIORATE THE PHENOMENON OF NON-SIZE-CONSTANCY IN HETEREOGENEOUS VR DISPLAYS.

JOSE DORADO, FRÉDÉRIC MERIENNE, PABLO FIGUEROA, JEAN-RÉMY CHARDONNET AND JOSÉ TIBERIO HÉRNANDEZ UNIVERSIDAD DE LOS ANDES (COLOMBIA) , ENSAM-PARITECH (FRANCE)

ABSTRACT.

The interception of the action-perception loop in virtual reality [VR] causes that understanding the effects of different display factors in spa tial perception becomes a challenge. For example, studies have reported that there is not size-constancy, the perceived size of an object does not remain constant as its distance increases. This phenomenon is closely related to the reports of underestimation of distances in VR, which causes remain unclear. Despite the efforts improving the spatial cues regarding display technology and computer graphics, some interest has started to focus on the human side. In this study, we propose a perceptual calibration method which can ameliorate the effects of non-size-constancy in heterogeneous VR displays. The method was validated in a perceptual matching experiment compar ing the performance between an HTC Vive HMD and a four-walls CAVE system. Results show that perceptual calibration based on interpupillary dis tance increments can solve partially the phenomenon of non-size-constancy in VR.
Introduction

The phenomenon of non-size-constancy is closely related to the enigma concerning distance perception in VR. Several studies have shown that distances are generally underestimated, but its causes remain unclear (see [23] for a complete review). With the advent of modern VR displays, the underestimation effects have begun to ameliorate but only when is measured using visually directed action tasks, such as blind walking ([6], [3]). In contrast, using other valid measurement tasks, such us blind-throwing and blind triangulated-pointing, the degree of underestimations increase and can vary significantly ([20]).

Similarly, the kind of display also influences the spatial perception. For example, divergences in large-immersive-projection displays [LIPDs] (such us immersive walls or CAVE systems) seem less stronger compared with head-mounted displays [HMDs]. Also, LIPDs have asymmetric performances which are closely related to the physical boundaries of the system. For example, there exists overestimation or underestima tion of distances depending on if the target object is located between the subject and the projection wall (negative stereoscopic parallax) or behind the wall (positive stereoscopic parallax). Hence, the perception of distance could be different, not only in comparison with the physical world but also between heterogeneous VR displays.

Due to its dependency on distance perception, the underestimation effects also impact the perception of size. In HMDs, objects tend to increase its apparent size whereas, in LIPDs, similar asymmetric effects are produced.

These effects were first reported by Kenyon et al. (2007) [7] using a CAVE system. They used a perceptual matching task to measure indirectly the perception of distance, requesting subjects to estimate the apparent size of a virtual object located at different distances over a virtual table, as it is shown in Figure 1. The virtual object was a replica of an equivalent physical one used as a reference. Besides, the distances were selected to represent the asymmetric effects related to the boundaries. Thus, their results showed that: only 55% of the population developed size-constancy and the errors were stronger for positive stereoscopic parallax as the target distance increased.

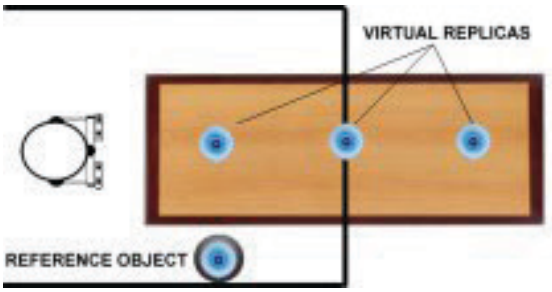


Fig. 1. Size-constancy-table experiment in a CAVE. Subject has to estimate the appar ent size of an object located at different distances using a physical object as reference.

In this study, we reproduced this experiment using modern technology and we explored the effect of perceptual calibration, which is a method that has shown good results improving distance perception. Also, we were interested in comparing the differences between heterogeneous VR displays.

For example, studying the same effects in HMDs is a challenge because it is impossible to visualize the physical object and its virtual replica at the same time. In this sense, we designed a perceptual calibration method targeting the differences in spatial performance between an HTC Vive HMD and a four-walls CAVE system. Therefore, the proposed method is an adaptation of Kenyon et al. experiment that is able to work in HMDs.

2. RELATED WORK

2.1 DISTANCE AND SIZE PERCEPTION IN THE REAL WORLD

Different methods have been proposed to assess how people perceive egocentric distances: (1) verbal estimates, the most straightforward method but also the less accurate [1]; (2) perceptual matching, where subjects are asked to reproduce a dis tance span based on a previously perceived physical target [25] and with a degree of underestimation relatively low at shorter distances [11]; (3) visually directed action, Title Suppressed Due to Excessive Length 3 the most popular method, where subjects are asked to estimate a distance per forming an equivalent action physically ([12], [24]). Hence, the size-constancy-table experiment is a kind of perceptual matching task.

Size perception is mostly visual and distance perception dependent. Is ruled by the phenomenon known as ”size-constancy” where an object is perceived as being of the same size regardless of its distance. The

perceive size of an object follows the size-distance-invariance hypothesis (SDIH) $s = d \tan(\alpha)$, with s the perceived size, d the distance to the object and α its visual angle. For this reason, the perception of size can also be used as an indirect measure of distance perception using perceptual matching tasks.

2.2 DISTANCE AND SIZE PERCEPTION IN VR

The most of the work on spatial perception in VR have been focused in egocentric distance perception, where a consensus exists that distances are generally underes timated, but its causes remain unclear. Most of past work were done using HMDs ([6], [26], [24]), where neither the limited field of view [FOV] nor the stereo viewing conditions nor the lack of realistic graphics contributes significantly to the underes timation effects.

Although these results suggest that the display factors are not the cause, recent evidence suggests that the causes could be related to the nature of the peripheral light stimulation induced by the display [9].
An interesting phenomenon is the differences that exist between HMDs and LIPDs, either using immersive walls

[21], [17]) or CAVE systems ([16], [13], [2]). We can highlight 3 important aspects about LIPDs on these studies:
(1) underestimation effects seem to be less stronger than in HMDs, (2) the physical space between the user and the projection screen is the most important factor and, (3) the effects are asymmetric with underestimation for objects at positive stereoscopic parallax and slight overestimation with objects at zero or negative stereoscopic parallax. Thus, as in the previous study, we took into consideration these aspects in our experiment.

Regarding size perception, it has been demonstrated that the SDIH holds in VR [18]. Also, when the complexity of the scene increases, subjects make better judg ments based on surrounding objects ([7], [16]). These characteristics make it suitable as a method to assess distance perception indirectly using perceptual matching ap proaches ([16], [4]). In this sense, the size-constancy-table test is an excellent task to study the phenomenon of non-size-constancy.

2.3 METHODS TO APPROXIMATE SPATIAL PERFORMANCE

Geometric calibration is focused on fixing possible discrepancies between the geomet ric FOV and the display FOV. Unfortunately, most of the work has been done in the field of augmented reality, where an object in the VE could be properly aligned with an object in the physical world [5]. Geometric calibration of Non-see-through HMDs is not straightforward, and there exists some evidence that even when the GFOV is calibrated, distance underestimation effects are reduced but not substantially [5].

Perceptual calibration is an alternative method that uses inverse geometric mod els to determine the perceptually correct viewing parameters ([15],[22]) and, its

4 NO AUTHOR GIVEN

application has shown to improve distance perception in CAVE systems [22]. The method is based on the idea of defining camera positions that are wider or deeper than the values obtained using standard calibration practices, with the purpose of influencing the perception of size and distance. We develop our first approach based on the study developed in [22], and we extended it to support HMDs.

Other methods use a ”computer vision” approach applying perspective projection adjustments: Minification

is a method which compresses the imaginary to artificially increasing the FOV of a display, and its application can improve distance perception ([8], [27], [10]). Lowering the horizon is another similar method which had good results ([14], [19]). Unfortunately, these methods have undesired effects on size perception due to the reduction on the imaginary or the apparent change in the perceiver's height.

PERCEPTUAL CALIBRATION METHOD

In VR, virtual cameras are usually calibrated via two viewing parameters: the inter pupillary distance [IPD] and the optical center depth [OCD] (the artificial middle point between the eyes). HMDs and LIPDs implement different immersion modali ties, but both share the same perceptual geometric model based on these parameters (Fig. 2). To produce the correct image impression in each eye, a diverged represen tation of the object is generated. Thus, the perceived object size *s* is proportional to its distance *d* and its visualization angle θ (Eq. 1). Likewise, the perceived object distance can be computed from the IPD and the convergence angle α (Eq. 1).

$$s = d \tan\theta \; 2 \qquad {}_{,d} = IP\,D \quad 2 * \tan {}^{\alpha}{}_{2} \quad (1)$$

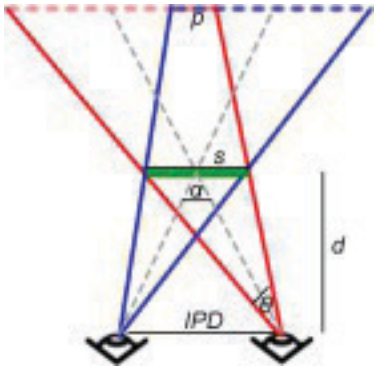


Fig. 2. Geometric model for the perception of size and distance of an object in VR.

Perceptual calibration is based on the idea that perception in VR is distorted, and defining calibration parameters based on the viewer's eyes morphology is a simplistic approximation, since it depends on how the actual geometry is interpreted by the Title Suppressed Due to Excessive Length 5 visual system which is out of observation ([22],[15]). A typical perceptual calibration procedure generates a set of perceptual camera positions by adjusting the IPD and OCD dynamically, with the purpose of finding a set of virtual cameras positions that approximate the subject's spatial perception to the physical world.

Based on the study developed in [22], we requested subjects to judge the super imposition of a virtual box with a physical one with the same dimensions. At first glance, the virtual box may seem slightly misaligned, bigger or smaller. Then, we request subjects to adjust the IPD and OCD dynamically using a standard Gamepad, until they consider the boxes perceptually match and we repeat the process iteratively at different target distances (0.5m, 1.5m, and 2.5m). The distances were selected to represent the different underestimation/overestimation effects of positive, close to zero or negative stereoscopic parallax (Logically, at 2.5m we assume the inverse value of 0.5m for the CAVE).

The proposed perceptual calibration procedure in two heterogeneous displays is shown in Figure (Fig. 3). In the CAVE, the physical box is made of glass which makes straightforward to fuse them and align them visually (Fig. 3 Left). In the HMD, we added a small screw at the top of the box, and we asked subject to compare its alignment by pointing to the same virtual position. A stick attached to the tracked HTC Vive controller enabled subjects to judge their alignment, comparing the visual cues with their proprioceptive

responses (Fig. 3 Right). In both kinds of displays, the box is located on the floor and at the same distances of the subject.

Fig. 3. Perceptual calibration procedure in both displays. Left, CAVE. Right, HMD.

In the HMD, we expected that subjects tend to set perceptual values for IPD and OCD that increase the perception of distance to compensate the underestimation effects. In the CAVE, due to the effects of negative stereoscopic parallax related to the physical boundaries, we expected that subjects tend to set perceptual values that decrease the perception of distance.

3.1 EFFECTS ON DISTANCE AND SIZE PERCEPTION

To validate our calibration procedure on distance and size perception, we used the size-constancy-table test. Similarly, we requested subjects to estimate the perceived size of an object located at 0.5m, 1.5m, and 2.5m over a set of virtual tables (Fig. 4). However, the set of tables was aligned starting with a physical one with the same dimensions, making a perceptual continuum. This is, the physical table provided a strong proprioceptive cue for distance and size perception, making subjects believe that its physical longitude was longer.

The experiment procedure was also similar: by using the physical object and the table as a reference, subjects had to estimate which could be the perceived size, either of a carbonated drink or a juice box, at the target distances and inside two heterogeneous displays (A HTC Vive HMD and a 4 walls CAVE).

They could adjust the scale of the virtual object according to the target distance using a standard Gamepad and, using only their sense of touch and their proprioceptive cues.

This restriction was important because subjects cannot compare the perceived size of the virtual object with the physical one in the HMD condition.

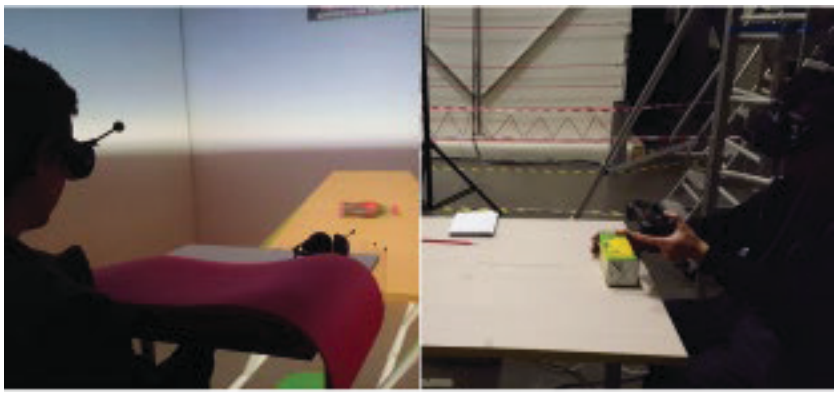


Fig. 4. Subject performing the size-constancy test. Left: CAVE, Right: HMD

3.2 HYPOTHESIS

Our hypothesis were as follows:

- H1. In the HMD condition, we expected greater IPD values or smaller OCD for all distances to compensate the
- underestimation effects.
- H2. In the CAVE condition, we expected smaller IPD values or greater OCD values to compensate the
- overestimation effect due to the negative stereoscopic parallax.
- H3. Although we did not expect similar perception of distance and size between displays, we expect size-constancy. In other words, the perceived size remains constant independently of distance for both displays.
- Title Suppressed Due to Excessive Length 7
- H4. There should exist a middle point where the spatial perception between both VR displays should converge.

3.3 PROCEDURE

Eight subjects (all male, M = 22.85 ± 1.06 years old) participated in our experi ment. All participants signed a letter of consent reporting a normal vision condition and good health at the moment of the experiment, without previous history of rel evant diseases.

We designed a within-subjects experiment where the participants perform the test with both displays with counterbalance order between subjects. In the beginning, we measured their IPD and calibrated the display according to these parameters. Also, to prevent bias with the first environment, we alternated the physical object between a carbonated drink and a juice box.

Subjects adjusted the perceived scale of the object 6 times at each target distance with aleatory order between trials. For each trial, the virtual object was presented with an exaggerate dimension equivalent to the 25% or 400% of its actual size.

Then, the subject had to adjust the scale of the object according to the perceived distance. In short, this gave us a configuration of 6 adjustments x 3 target distances x 2 objects x 2 VR display conditions.

3.4 RESULTS

A paired sample t-test was used for all the measures. Figure 5 Left shows the results of the perceptually calibrated viewing parameters defined for each display. In the HMD condition, contrary to our predictions, we found significant greater IPD values (M = 0.0311 ± 0.022), $t(7)$ = 0.347 compared with the average subject physical IPD. In term of OCD values, we did not found significant differences (M = 0.051 ± 0.056), $t(7)$ = 0.333 in comparison with the default OCD. These results suggest that H1 partially holds for the IPD parameter.

Contrary to our expectations, in the CAVE condition we found significant greater IPD values (M = 0.081±0.054), $t(7)$ = 0.347 compared with the average subject physical IPD. Also, we did not found significant differences with in terms of OCD values M = 0.055 ± 0.013), $t(7)$ = 0.333 with a greater variability, which suggest that H2 did not hold.

We believe that these partial successfully results could be due to artifacts in our calibration procedure. In the CAVE condition, we believe this could be related to a problem of perspective: from the viewer's point of view, the virtual box seems misaligned to the physical, even when its size was the same.

In the HMD condition, it was difficult for the subjects to judge the alignment of the box using only propri oceptive cues and, to compare the declination angle of the stick because we did not provide rigid body collision feedback.

Hence, we noticed that some subjects were not confident enough about the perceptual alignment of the boxes.

Although we did not confirm completely hypothesis H1 and H2, we got some interesting results on the effects in distance and size perception based on the size constancy-table test, which are described in Figure 5 Right. Using a paired sample t-test, we found that the spatial perception is significantly different between the CAVE and the HMD at 0.5m (M = 1.28 ± 0.10, M = 1.15 ± 0.10, $t(7)$ = 0.01),

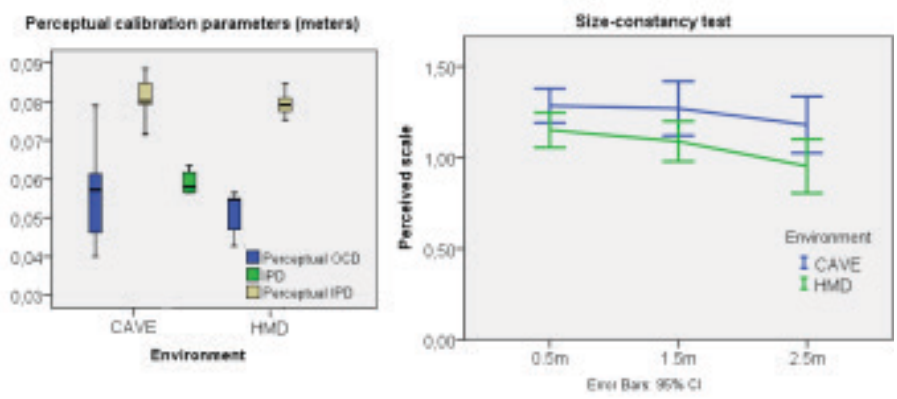


Fig. 5. Left: Result of the perceptual calibration procedure. Right: Results of the size constancy-table test in both displays.

at 1.5m (M = 1.26 ± 0.16, M = 1.08 ± 0.11, $t(7)$ = 0.023) and at 2.5m (M = 1.18 ± 0.16, M = 0.95 ± 0.16, $t(7)$ = 0.014). Also, we noticed a great tendency to size-constancy in the CAVE condition and non-size-constancy in the HMD condition, which suggest that H3 partially holds for the CAVE. Finally, due to the lines did not intercept, this means that H4 did not hold, indicating that the perception of size is diametrically different.

4 DISCUSSION

We proposed a novel method to ameliorate the phenomenon of non-size-constancy in heterogeneous VR displays. Regarding the perceptual calibration procedure, ad justing the IPD was more effective than adjusting the OCD because it produces stronger effects on the object's projected image. Interestingly and contrarily to our predictions, subjects rather perform IPD increments in the CAVE condition than decrements, causing an equivalent increase in the perceived size.

We can explain this anomaly as consequence of the different sensory modalities used to estimate the size of the bottle (visual vs. proprioceptive). However, independently of this effect, there is a higher tendency in the CAVE to size-constancy for negative stereoscopic parallax (at 0.5m and 1.5 m) and very few underestimation effects for positive stereo scopic parallax (at 2.5m), which confirms the influence of the physical boundaries reported by Kenyon et al.

Unfortunately, in HMDs we did not get size-constancy, the perceived distances were underestimated and the perceived size varies as consequence. There may exist some explanation for these poor results such as the use of different sensory modal ities or the deficiencies in our calibration procedure. However, we believed that the underestimation effects are so strong in HMDs that just adjusting the IPD is not enough because the range of possible values is limited by the threshold were stereop sis is comfortable. Also, the HMD condition lacks other important spatial cues, such as visualizing of the physical body or being inside a familiar environment. In this

sense, we consider that the scope of perceptual calibration based on IPD increments was limited for HMDs.

5 FUTURE WORK

The main challenge in the perceptual calibration procedure for HMDs is the impos sibility of visualizing simultaneously the physical object and its virtual replica. As future work, we are exploring ways to perform perceptual matching tasks that do not require seeing the physical world but requires some action. A great part of the popularity of visually directed action methods, such as blind walking, is their inde pendence on the visual stimulus during the measurement procedure. For example, a possible method could be using subject's affordances to estimated the perceived size of a gap at different distances using their body lenght as reference [4].

Finally, we are planning to include other spatial cues, such as allowing subjects to visualize their hands (using a Leap Motion) and providing a familiar environ ment recreating the test room in the VE. Also, due to limitations of using IPD increments, we are planning to explore the methods based on ”computer vision” ap proaches described in section 2.3 that performs perspective projection adjustments. Our objective is to find ways that we can alter the image formation process produc ing images that have a sense (from the perception point of view) and can influence the perception of size and distance.

REFERENCES

1. Andre, J., Rogers, S.: Using verbal and blind-walking distance estimates to investigate the two visual systems hypothesis. Attention, Perception, & Psychophysics **68**(3), 353– 361 (2006)
2. Bruder, G., Argelaguet, F., Olivier, A.H., L´ecuyer, A.: Cave size matters: Effects of screen distance and parallax on distance estimation in large immersive display setups. Presence: Teleoperators and Virtual Environments **25**(1), 1–16 (2016)
3. Creem-Regehr, S.H., Stefanucci, J.K., Thompson, W.B., Nash, N., McCardell, M.: Egocentric distance perception in the
4. oculus rift (dk2). In: Proceedings of the ACM SIGGRAPH Symposium on Applied Perception. pp. 47–50. ACM (2015)
5. Geuss, M., Stefanucci, J., Creem-Regehr, S., Thompson, W.B.: Can i pass?: using affordances to measure perceived size in virtual environments. In: Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization. pp. 61–64. ACM (2010)
6. Kellner, F., Bolte, B., Bruder, G., Rautenberg, U., Steinicke, F., Lappe, M., Koch, R.: Geometric calibration of head-mounted displays and its effects on distance estimation. IEEE transactions on visualization and computer graphics **18**(4), 589–596 (2012)
7. Kelly, J.W., Cherep, L.A., Siegel, Z.D.: Perceived space in the htc vive. ACM Trans actions on Applied Perception (TAP) **15**(1), 2 (2017)
8. Kenyon, R.V., Sandin, D., Smith, R.C., Pawlicki, R., Defanti, T.: Size-constancy in the cave. Presence: Teleoperators and Virtual Environments **16**(2), 172–187 (2007) 8. Kuhl, S.A., Thompson, W.B., Creem-Regehr, S.H.: Minification influences spatial judg ments in virtual environments. In: Proceedings of the 3rd symposium on Applied per ception in graphics and visualization. pp. 15–19. ACM (2006)
9. 10 No Author Given
10. Li, B., Nordman, A., Walker, J., Kuhl, S.A.: The effects of artificially reduced field of view and peripheral frame stimulation on distance judgments in hmds. In: Proceedings of the ACM Symposium on Applied Perception. pp. 53–56. ACM (2016)
11. Li, B., Zhang, R., Kuhl, S.: Minication affects action-based distance judgments in oculus rift hmds. In: Proceedings of the ACM Symposium on Applied Perception. pp. 91–94. ACM (2014)
12. Loomis, J.M., Da Silva, J.A., Fujita, N., Fukusima, S.S.: Visual space perception and visually directed action. J. of Experimental Psychology: Human Perception and Per formance **18**(4), 906 (1992)
13. Loomis, J.M., Knapp, J.M.: Visual perception of egocentric distance in real and virtual environments. Virtual and adaptive environments **11**, 21–46 (2003)
14. Marsh, W.E., Chardonnet, J.R., Merienne, F.: Virtual distance estimation in a cave. In: International Conference on Spatial Cognition. pp. 354–369. Springer (2014) 14. Messing, R., Durgin, F.H.: Distance perception and the visual horizon in head-mounted displays. ACM Transactions on Applied Perception (TAP) **2**(3), 234–250 (2005) 15. Mestre, D.R.: Perceptual calibration in virtual reality applications. Electronic Imaging **2016**(4), 1–6 (2016)
15. 16. Murgia, A., Sharkey, P.M., et al.: Estimation of distances in virtual environments using size constancy. The International Journal of Virtual Reality **8**(1), 67–74 (2009) 17. Naceri, A., Chellali, R., Dionnet, F., Toma, S.: Depth perception within virtual environ ments: comparison between two display technologies. International Journ. on Advances in Intelligent Systems **3** (2010)
16. Nakamizo, S., Imamura, M.: Verification of emmert's law in actual and virtual en vironments. Journal of
17. physiological anthropology and applied human science **23**(6), 325–329 (2004)
18. Ooi, T.L., Wu, B., He, Z.J.: Distance determined by the angular declination below the horizon. Nature **414**(6860), 197 (2001)
19. Peer, A., Ponto, K.: Evaluating perceived distance measures in room-scale spaces using consumer-grade head mounted displays. In: 3D User Interfaces (3DUI), 2017 IEEE Symposium on. pp. 83–86. IEEE (2017)

20. Plumert, J.M., Kearney, J.K., Cremer, J.F., Recker, K.: Distance perception in real and virtual environments. ACM Transactions on Applied Perception (TAP) **2**(3), 216–233 (2005)
21. Ponto, K., Gleicher, M., Radwin, R.G., Shin, H.J.: Perceptual calibration for immersive display environments. IEEE Trans. on visu. and computer graphics **19**(4), 691–700 (2013)
22. Renner, R., Velichkovsky, B., Helmert, J.: The perception of egocentric distances in virtual environments. ACM Comput. Surv.(to appear, 2014) Google Scholar 24. Sahm, C.S., Creem-Regehr, S.H., Thompson, W.B., Willemsen, P.: Throwing versus walking as indicators of distance perception in similar real and virtual environments. ACM Transactions on Applied Perception (TAP) **2**(1), 35–45 (2005) 25. Sinai, M.J., Krebs, W.K., Darken, R.P., Rowland, J., McCarley, J.: Egocentric distance perception in a virutal environment using a perceptual matching task. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 43, pp. 1256–1260. SAGE Publications Sage CA: Los Angeles, CA (1999)
23. Willemsen, P., Gooch, A.A., Thompson, W.B., Creem-Regehr, S.H.: Effects of stereo viewing conditions on distance perception in virtual environments. Presence: Teleop erators and Virtual Environments **17**(1), 91–101 (2008)
24. Zhang, R., Nordman, A., Walker, J., Kuhl, S.A.: Minification affects verbal-and action based distance judgments differently in head-mounted displays. ACM Transactions on Applied Perception (TAP) **9**(3), 14 (2012)

NAMED ENTITY RECOGNITION USING NEURAL NETWORKS FOR CLINICAL NOTES

EDSON FLOREZ1, FREDERIC PRECIOSO1, ROMARIC PIGHETTI2 AND MICHEL RIVEILL1

1 UNIVERSITÉ CÔTE D'AZUR, NICE 06000, FRANCE

2 FRANCE LABS, SAINT-LAURENT-DU-VAR 06700, FRANCE

FLOREZ@I3S.UNICE.FR

ABSTRACT.

Currently, the best performance for Named Entity Recognition in med-ical notes is obtained by systems based on neural networks. These supervised systems require precise features in order to learn well fitted models from training data, for the purpose of recognizing medical entities like medication and Adverse Drug Events (ADE). Because it is an important issue before training the neural network, we focus our work on building comprehensive word representations (the input of the neural network), using character-based word representations and word representations. The proposed representation improves the performance of the baseline LSTM. However, it does not reach the performances of the top per-forming contenders in the challenge for detecting medical entities from clinical notes [13].

Keywords: Named Entity Recognition, Clinical Notes, Adverse Drug Events, Deep Learning, LSTM.

INTRODUCTION

Patients are often subject to multiple treatments, which may be the cause of adverse effects. Therefore, it is necessary to establish if an Adverse Drug Event (ADE) has occurred after taking medicines. ADE refers to any adverse event occurring at the time a drug is used, whether it is identified as a cause of the event or not. In case one can establish a relation between the ADE and the drug, then the relation is considered as an Adverse Drug Reaction (ADR), which is a Relation Extraction (RE) task.

For the purpose of identifying ADE mentions, we use medical notes provided in EHR (Electronic Health Records). These notes contain mentions of medical entities like medications, ADE (Adverse Drug Event) and symptoms. These terms have to be identified and classified in the right category. This classification problem is known as Named Entity Recognition (NER). A named entity is a term (one or many words) that can be annotated with a label (tag) if it belongs to some category.

Named Entity Recognition has been performed with systems based on Machine Learning and Deep Learning algorithms, for automatic detection of entity mentions in medical notes, entities such as ADEs, indications, medication name and its attributes (dosage, frequency, route, duration) [13]. In this work neural networks are used for

NER in clinical notes, using several word representation together to improve the per-formance. Section present some related works in the domain of NER. The several fea-tures used as well as the network used

are explained in Section 2. In Section 3 the models performance using the dataset provided by the MADE1.0 Challenge [13] are presented.

1. RELATED WORK

Conditional Random Fields (CRFs) is a machine learning algorithm used for ADR ex-traction [11], CRF can take context (around the current word) into account for sequence modeling, it takes every neighbour word in a fixed window of words [3].

Other Machine Learning algorithms like Support Vector Machines (SVMs) are commonly used for NER. Gurulingappa et. al. [7] built a system for the identification and extraction of potential adverse events of drugs with SVM.

Their dataset is an ADE corpus from MEDLINE (Medical Literature Analysis and Retrieval System Online) case reports that are manually annotated. The corpus contains annotations for the mentions of drugs, ADE, and relations between drugs and medical conditions representing clear adverse reactions (relation drug-cause-condition).

The CLEF (Cross-Language Evaluation Forum) eHealth Evaluation Lab provides system performance for NER (Task 1b in CLEF 2015 [10]) using the QUAERO French Medical Corpus [9]. It has ten categories for annotations of medical entities, with data collected from the EMEA (European Medicines Agency) documents and titles of re-search articles indexed in the MEDLINE database.

A Dictionary-based concept recog-nition system overcame CRF and SVM classifiers in CLEF 2015 [10] on the MEDLINE corpus, according to the Exact Match metric, which considers a term (word or group of words that have a label) as correctly classified only if all the words in the term received the correct label.

Deep learning models like CNN (Convolutional Neural Network) are used to detect the presence of ADR [4], such as in binary classification problem on two datasets (from Twitter and case reports [7]). Overall, CNN appears to perform better compared to other more complex CNN variants that have a RNN (Recurrent Neural Network) layer [7]. However, CCNA (Convolutional Neural Network with Attention) is better on the da-taset of case reports. Overall, results on the case reports are better than those on the Twitter dataset. Tweets contain a lot of ill-grammatical sentences and short forms [4] that hinders the performances, which highlights the importance of de-noising the data.

The adverse event detection problem focused on clinical notes is a sequential prob-lem, and RNN models are specialized for it because at time step t, the recurrent node takes as input the outputs produced by the previous state. RNN models were limited to make separate classifications at every time step on an input sequence [2], but another RNN architecture, known as Long Short-Term Memory (LSTM), is designed to take into account the long-time dependencies between relevant input and targets.

LSTM was applied to sequential problems such as Handwriting Recognition [2] and Named Entity Recognition [8]. LSTM exploits the long term label dependencies for sequence labelling in clinical text, e.g. in the sentence “the patient has internal bleeding (ADR) sec-ondary to warfarin (Medication)”, the label for ADR is strongly related to the label prediction of Medication, then Warfarin is labelled as Medication using information of previous ADR tag (internal bleeding), which is stored in the memory of LSTM cells.

LSTM was used with an annotated corpus of English Electronic Health Records (EHR) from cancer patients in [1], with labels for several medical entities (like Adverse Drug Event (ADE), drug name, dosage) and

relations between entities. The best LSTM version in [1] is the Approximate Skip Chain CRF-RNN network, which implements a CRF algorithm after the bidirectional LSTM output. This network has a high precision for DrugName detection, but a low precision for ADE, probably because the dataset in unbalanced and has less ADE samples.

Results of NER algorithms dedicated to ADE detection are collected in the review article [6]. This review shows that Machine learning and Deep Learning algorithms are outstanding at this task. However, the performance presented in this review were ob-tained on different datasets, making the comparison somewhat unfair. Comparing the best result reported in [4] and [7] using the same dataset (last lines of Table 1), one can observe that Gurulingappa et al. [7] obtained slightly better results on Recall, Precision and F-score.

Table 1. Methods for ADE extraction

Study	Ref.	Method	Size	Recall	Prec.	F1
Nikfarjam	[5]	Lexical pattern-matching	1200	0.66	0.70	0.68
Nikfarjam	[3]	Supervised learning via Conditional Random Fields (CRFs)	1559	0.78	0.86	0.82
Jagannatha	[1]	Bi-LSTM-CRF (Skip-CRF-Approx.)	1154	0.83	0.81	0.82
Huynh	[4]*	CNNA (Convolutional Neural Network with Attention)	2972	0.84	0.82	0.83
Gurulingappa	[7]*	SVM (Support Vector Machines)	2972	**0.86**	**0.89**	**0.87**

*Systems using the same dataset

LSTM model has shown to be appropriate on the state of the art for sequential problems. However, in order to improve performance, it is important to feed the network with an appropriate input representation (an embedding) [12]. This representation replaces each unique word with a dense vector representation, which tries to provide closer vectors among word synonyms or related words.

In [1] the embedding layer values used were initialized using a Skip-gram word embedding. The Skip-gram embedding was calcu-lated using unlabelled data from PubMed open access articles, English Wikipedia and an unlabelled EHR corpus. We can also improve the precision of LSTM with additional features for its input, such as character-level features from each word extracted using CNN or LSTM [11], and then concatenate character and word representations inspired by the work of Chiu et. al. [12]. All this was implemented in our work, as described in the following section.

2 MODEL

In our final model, we use a comprehensive word representation, which concatenates character-level representations, word embedding and POS features. This is described in the following subsections, as well as the full network using that representation to solve the NER task.

2.1 FEATURES

The character-level features can exploit prefix and suffix information about words [17], to have closer representations among words of the same category. This is particularly useful for terms that may be Out-Of-Vocabulary (appearing in the test data and not in the training data). OOV is a common issue with domain specific words, and prefix and suffix representations can help a lot. For example, the words “Clonazepam”

and “Lo-razepam” both belong to the medication category in the medical context and may be OOV. However they share the same suffix, making them closer to each other on a char-acter-level feature. Therefore we build a LSTM network (see sub-section 2.2) that get representations of words based on their characters.

Another feature used is Part-of-speech (POS), which tags the words with labels like noun, verb, adjective, adverb, etc. It classifies words according to its roles within the grammatical structure of the sentence. Medications for example will always belong to the Noun category, making them close together with respect to this feature. The tagging was performed using an Averaged Perceptron algorithm [15].

Finally, we also use word embeddings learned from a large corpus, to consider the contexts in which words appear usually. It can create similar vectors (representations) for words that appear in similar contexts, such as the names of different countries. The word embedding of dimension 200 provided by [1] were used, as well as another of 300 dimensions provided by FastText [14]. Both are pre-trained with skip-gram using unlabeled data mainly from Wikipedia.

2.2 NETWORK DESCRIPTION

Long Short-term Memory Networks (LSTMs) can learn long term dependencies among the words in the sentence [1]. LSTM keeps information in a memory-cell that is updated using input i_t and forget gates f_t [17].

The character-level embedding for words was built by a Bi-LSTM network (repre-sented on the bottom left of Figure 1). First, each character takes an integer value from a lookup table, then it is replaced by a one-hot vector.

The final state of the forward and backward LSTM is the representation of the suffix and prefix of the word. The Char-acter-level embedding is the concatenation of both LSTM layers, so with LSTM layers of 20 cells (units), we get a vector of 40 dimensions. This character-level representation is concatenated to the word embedding and the POS feature to form the final compre-hensive word representation (see Fig. 1) [17].

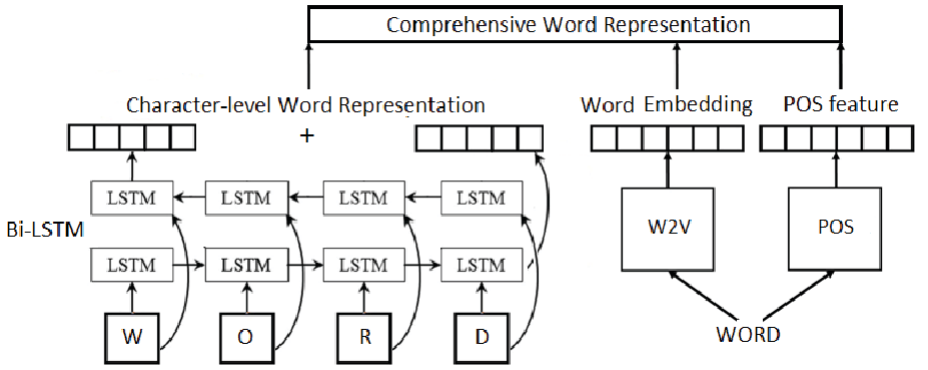


Fig. 1. Comprehensive word representation

The comprehensive word embedding is the input of a Bi-LSTM network, which takes a sequence of words and returns a sequence of hidden states at every time step (see Fig. 2).

The raw sentence is processed with a regular expression tokenizer into sequence of tokens. Sentences longer than the sequence length were cropped to size, and shorter sentences were pre-padded with masks.

The forward and backward LSTM layers get hidden state sequences, which represent the left and right context of the sen-tence at every time step (word), and their concatenation is the representation of a word in context [16].

The bidirectional LSTM provides scores for every possible label for each word, its out-put (hidden states) feed the inference layer for tagging each word independently (see Fig. 2). For that, the hidden states are connected by a dense layer (i.e. fully connected layer) to each possible label, and a Softmax function over the score of all possible labels produces a probability for each label (values between 0 and 1 that together sum 1), which is used to get the predicted label.

The predictions (labels probabilities) of the Softmax output is evaluated with the correct class (true labels). The target labels consist in an integer vector where each element represents the position of the number 1 in a one-hot encoding. Categorical cross-entropy is the loss function used, which penalizes the deviation between the predicted and target (true) labels during training. Then the optimization function will minimize the loss of the correct labels sequence.

For training the network, the input and output of the network will be the sequence of words (each word replaced by its comprehensive word representation) and its corre-sponding labels (see Fig. 2), and LSTM will try to learn a model that minimize the error of the predicted label.

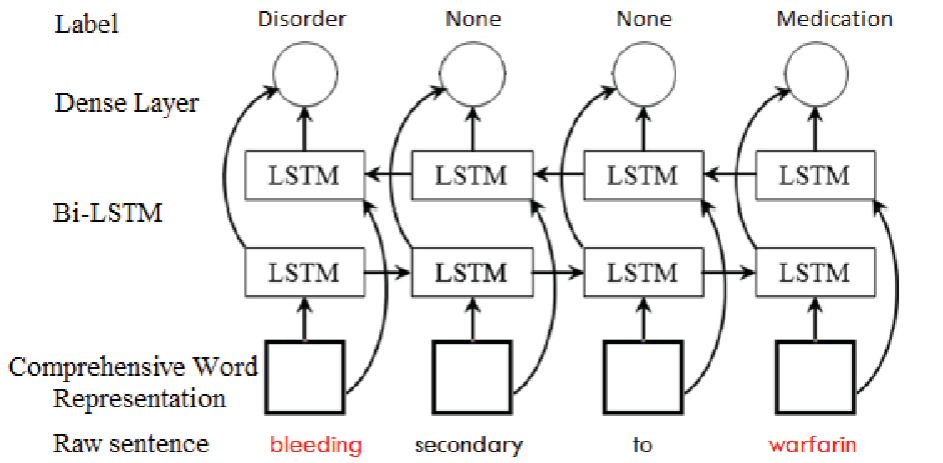


Fig. 2. LSTM network for tagging

3 RESULTS AND DISCUSSION

The dataset for our experiments was provided by the MADE1.0 Challenge [13]. It was created with 1092 EHR notes from 21 cancer patients [1]. It contains annotations of ADEs, indications, other signs and symptoms, medication, dosage, route, frequency, duration, severity.

These annotations are used in the Named entity recognition (NER) task, and the dataset also has relations among those medical entities for the Relation Extraction task, like the relation Adverse between Medication and ADE annotations. In the NER task, we the goal is to identify and annotate the medical entities found in the raw clinical notes.

The models were compared with the same parameters and training dataset as those of the MADE challenge. The dataset is split into training (80% of the data) and testing (20 of the data). The results are

shown in Table 2, with results for models with random initialization of word vectors (baseline), W2V of 200 dim [1], W2V(FT) FastText of 300 dim [14], POS features (46 tags) and Character-level word representation Char(LSTM) of length 40.

We improved the performance using the word embedding of FastText (W2V(FT)) more than using the one of W2V[1]: FastText (W2V(FT)) got about 0.22 more in F1 than W2V[1]. We observed the highest improvement over the baseline (randomly ini-tialized model) with character-level representations and POS tags together, it increases the F1 of about 0.2. W2V(FT) only with the Char(LSTM) provides a small increase in F1, while POS alone does not increase anything.

Table 2. Performances of models for NER

Model	Recall	Precision	F1
Baseline	0,686	0,704	0,695
W2V[1]	0,668	0,689	0,678
Char(LSTM) + POS	0,659	0,678	0,668
W2V(FT)	0,694	0,721	0,707
W2V(FT) + POS	0,691	0,719	0,704
W2V(FT) + Char(LSTM)	0,692	**0,724**	0,708
W2V(FT) + Char(LSTM) + POS	**0,700**	0,721	**0,710**

Note: Parameters batch size 32, sequence length 60, 100 LSTM cells, learning rate 0.1

The best model (W2V+Char(LSTM)+POS) was trained with 100% of the training files, then it created the predicted annotations for the test dataset of the MADE Challenge. Table 3 shows the official results validated by the MADE challenge [13], the best result of 2 runs for standard (W2V [1]) and extended evaluation (W2V(FT)).

The usage of more hidden units (200 or 300 LSTM cells) did not significantly influenced the model performance, and big values (60, 70, 80) of the sequence length (number of words by sequence) gave better results in our experiments with the clinical notes of MADE da-taset. The most appropriate initial value for the learning rate was 0.1, a smaller learning rate decreased the performance and increased the running time. The results are good but an additional strategy is still necessary to reach top performance systems (the best has 0.829 in F1 [13]). An additional layer of conditional random fields used over the output of LSTM (in the tagging layer), which take into account the dependencies be-tween labels to get an accurate score like in [1] would be interesting to test.

Table 3. Performances of models for NER task in MADE Challenge

Model	Recall	Precision	F1
W2V[1] + Char(LSTM) + POS	0,720	0,681	0,700
W2V(FT) + Char(LSTM) + POS	**0,748**	**0,716**	**0,732**

CONCLUSIONS

We implemented a LSTM network to solve the named entity recognition problem found on the Adverse Drug Reaction detection. This neural network requires good input fea-tures for training, so we built character-level features extracted with another LSTM, that were used in conjunction with word representations as a comprehensive word rep-resentation. This conjunction of features increased the performance of the LSTM,

but it does not allow the LSTM alone to reach the best performance achieved for the task. Therefore, as future work, investigating the use of an additional technique for the net-work, as the Attentional model for RNN that gives more weight to words that are more important, sounds promising. For the final purpose of identify which treatments may be the cause of an Adverse Drug Event (ADE), known as the Adverse Drug Reaction (ADR) relation, we will build a full system for NER and Relation Extraction.

REFERENCES

1. Jagannatha, A. N., Yu, H.: Structured prediction models for RNN based sequence labeling in clinical text. Proceedings of the Conference on Empirical Methods in Natural Language Processing (2016).
2. Liwicki, M., Graves, A., Fernàndez, S., Bunke, H., Schmidhuber, J.: A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. Proc. 9th International Conference on Document Analysis and Recognition, vol. 1 (2007).
3. Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., Gonzalez, G.: Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association (2015).
4. Huynh, T., He, Y., Willis, A., Rüger, S.: Adverse drug reaction classification with deep neural networks. pp. 877-887 (2016).
5. Nikfarjam, A., Gonzalez, G.: Pattern mining for extraction of mentions of adverse drug re-actions from user comments. In: Proceedings of the American medical informatics associa-tion (AMIA) annual symposium, pp. 1019–26 (2011).
6. Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G.: Utilizing social media data for pharmacovigilance: A review. Journal of bio-medical informatics 54 (2015).
7. Gurulingappa, H., Mateen-Rajpu, A., Toldo, L.: Extraction of potential adverse drug events from medical case reports. Journal of biomedical semantics 3(1), (2012).
8. Jagannatha, A., Yu, H.: Bidirectional rnn for medical event detection in electronic health records. Proceedings of the conference of Association for Computational Linguistics. North American Chapter. Meeting, vol. 2016. NIH Public Access (2016).
9. Neveol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The QUAERO French Med-ical Corpus: A Resource for Medical Entity Recognition and Normalization. Fourth Work-shop on Building and Evaluating Resources for Health and Biomedical Text Processing - BioTxtM2014, pp. 24-30 (2014).
10. Neveol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P.: Task 1b of the CLEF eHealth Evaluation Lab 2015: Clinical Named Entity Recognition. CLEF 2015 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2015).
11. Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., Xu, H.: Entity recognition from clinical texts via recurrent neural network. BMC medical informatics and decision making, 17(2), 67 (2017).
12. Chiu, J., and Nichols, E.: Named entity recognition with bidirectional lstm-cnns. arXiv pre-print arXiv:1511.08308 (2015).
13. Yu, H., Jagannatha, A., Liu, F., Liu, W.: NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (https://bio-nlp.org/index.php/announcements/39-nlp-challenges) (2018).
14. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016).
15. Honnibal, M., NLTK Library. https://www.nltk.org/api/nltk.tag.html, last accessed 2018/03/10
16. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM networks. In Proceedings of International Joint Conference on Neural Networks, vol. 4, pp. 2047-2052 (2005).
17. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016).

VACAPP – POUR SUIVRE LE SCHÉMA DE VACCINATION DES ENFANTS ET DES ADULTES

CÉSAR MUÑOZ1 ET HELGA DUARTE2

1 SPÉCIALISATION EN GOUVERNEMENT EN LIGNE - GEL UNIVERSIDAD NACIONAL DE COLOMBIA – SEDE BOGOTÁ, COLOMBIA 2 UNIVERSIDAD NACIONAL DE COLOMBIA – SEDE BOGOTÁ, COLOMBIA FACULTÉ DE GÉNIE

{CEMUNOZB,HDUARTE}@UNAL.EDU.CO

RÉSUMÉ.

Ce projet présente un outil technologique, VacApp, qui aide la popula-tion à surveiller, contrôler et gérer son propre programme de vaccination enre-gistré dans le cadre du Programme Elargi de Vaccination (PAI), conformément aux directrices de vaccination établies par l'Organisation Panaméricaine de la Santé, l'Organisation Mondiale de la Santé - OMS, et par le Département de Santé de la ville de Bogotá, tout en tenant compte du cadre de travail du Gouver-nement en Ligne fourni par le ministère des TIC. Grâce à cet outil technologique, les citoyens peuvent vérifier l'état actuel de leur schéma de vaccination, recevoir des alertes pour compléter ce programme et se renseigner sur les journées de vaccination actives (qui comprennent le lieu et la date de vaccination). En outre, le programme peut envoyer des informations afin de sensibiliser et d'éduquer la population sur l'importance de s'engager à un programme de vaccination

Mots clés - Vaccination, Santé Publique, TIC, Ministère de la Santé, Pro-gramme Elargi de Vaccination - PAI, Technologie de la Santé.

1 INTRODUCTION

Le Programme Elargi de Vaccination (PAI) [1] vise à éliminer, éradiquer et contrôler les maladies à prévention vaccinale en Colombie, principalement chez les moins de 5 ans, afin de réduire les taux de mortalité et de morbidité causés par ces maladies.

La ville de Bogotá compte 7 millions de personnes inscrites dans la base de données du PAI où il y a le registre de plus de 40 millions de vaccins appliquées.

Malheureusement, cette population ne dispose pas d'un outil lui permettant de vérifier son statut vaccinal actuel, de telle sorte qu'il l'aide à compléter le programme si certains vaccins leur manquent. Ou tout simplement de recevoir des informations relatives aux systèmes de vaccination. Le but de ce projet est de proposer une application mobile appelée VacApp qui permet, au début, de satisfaire ce besoin (en outre de celles dictées dans [3, 4 y 5]). Cette application peut également avoir des fonctionnalités étendues au service du secteur de la santé.

Cette proposition renforce le concept de société de l'information et de la connaissance, approche innovante des Technologies de l'Information et Communications (TIC) [2], car il facilite le transfert

d'information, la gestion et le contrôle des données et l'utilisa-tion et l'appropriation des connaissances par les citoyens. Tout cela en transformant les processus et les rendre plus efficaces pour les utilisateurs, ainsi que pour les entités fournissant des services de vaccination.

2 ETAT DE L'ART

Depuis plusieurs années, le PAI [7] est responsable de la planification, de l'organisation, de la coordination, de l'exécution, du suivi, du contrôle et de l'évaluation des processus et procédures de vaccination afin de réduire les taux de mortalité et de morbidité [9].
Pour développer ce projet, nous avons tenu compte du PAI du Département de Santé de la ville de Bogotá et des recherches qui ont été menées dans d'autres instances de santé de différentes villes de la Colombie. Au même temps, nous avons fait une explo-ration sur le développement et l'utilisation d'applications mobiles dans le domaine de la santé, en particulier dans le domaine de la vaccination, dans différents pays d'Amérique latine et de l'Espagne.

2.1 BOGOTÁ

À Bogotá, nous avons eu accès au PAI du Département de la Santé, et nous avons connu tous les processus et procédures mis en œuvre par cette entité. L'objectif général de ce programme est de mener des actions de planification, d'organisation, de coordination, d'exécution, de suivi, de contrôle et d'évaluation des processus et des procédures afin d'atteindre la couverture vaccinale dans la ville.

Parmi les objectifs spécifiques, nous pouvons mentionner i) Maintenir le programme PAI comme une politique de santé publique prioritaire ii) Reconnaître les caractéris-tiques de l'environnement changeant dans lequel le PAI est développé. iii) Identifier les faiblesses existantes dans le PAI et appliquer la gestion stratégique afin d'obtenir des résultats favorables.

D'autre part, les stratégies et tactiques de vaccination à Bogotá ont comme objectif principal de réaliser la conformité des schémas de vaccination dans la population de la ville. Ses objectifs spécifiques sont : i) Réduire les causes des "occasions manquées" en vaccination ii) Surveiller la population pour initier, poursuivre et compléter le pro-gramme de vaccination iii) Sensibiliser et promouvoir le respect des services de vacci-nation aux parents et/ou soignants.

STRATÉGIES DE VACCINATION DU PAI.

À Bogotá, le PAI envisage l'utilisation de trois stratégies fondamentales pour offrir le service de vaccination à la population cible :

1. Actions permanentes de vaccination : toutes les actions qui se déroulent dans le PAI, 365 jours par an, pour l'application de toutes les vaccines du schéma na-tional au niveau institutionnel via des postes fixes, à domicile, ou à travers des équipes mobiles.

2. Journées de vaccination intensive : activités intensives telles que Conférences et journées avec mobilisation massive de la population en un jour ou en une courte période de temps, afin d'appliquer le plus grand nombre de doses de vaccins.

3. Nouvelles actions de vaccination : intensification de la vaccination à domicile de la population à risque. Le but est d'interrompre la transmission d'une maladie dans une courte période de temps.

SYSTÈME D'INFORMATION DU PAI

Dans le cadre du PAI - Bogotá, le composant système d'information est chargée de promouvoir l'échange d'informations entre les différentes entités participants au pro-cessus de vaccination.

Pour cette raison, il a été créé le Système d'information PAI qui a été développé pour traiter l'information générées dans les services de vaccination des établissements de santé de manière rapide et efficace.

Le système web PAI version 2.0 a été créé par la nécessité de mettre à jour l'application web existante - PAI version 1.0. (Développé dans Visual Studio 2003), où le processus de développement a été complété avec la correction des fonctionnalités, tandis que de nouvelles fonctionnalités ont été ajoutées. De même, les rapports sont ajustés confor-mément aux réglementations en vigueur du Ministère de la Santé et de la Protection Sociale.

L'infrastructure des systèmes d'information est gérée par la Direction de TIC du Mi-nistère de la Santé [8]. La figure suivante montre les principaux composants du système d'information.

Fig. 1. Diagramme du système d'information PAI [8]

En tant que caractéristique principale, le système utilise une base de données qui enregistre en temps réel les vaccins appliqués par chacune des entités sanitaires ratta-chées au Secrétariat de Santé du District (ou tout simplement Secrétariat de Santé de Bogotá).

En outre, la ville de Bogotá a réussi la cohésion de toutes les entités de santé, telles que les l'EAPB (Entités Administrant les Régimes de Bénéfices de Santé, Entités Promo-trices de la Santé du Régime Contributif, Médecine prépayée et Ambulance Prépayé), EPS/IPS (Entités/Institutions Prestataires de soins de Santé obligatoire), etc., ce que lui permet d'avoir un système d'information nominal unique au niveau de la ville, géré par le Secrétariat de Santé de Bogotá, ce qui garantit un gouvernement et une gouvernance de l'information concernant la vaccination.

2.2 LE MINISTÈRE DE LA SANTÉ ET DE LA PROTECTION SOCIALE ET LE RESTE DES VILLES DE LA COLOMBIE

Contrairement à Bogotá, aucune autre ville de Colombie ne dispose d'un système d'information sur la vaccination tel qu'il est décrit dans la section précédente.

Les autres villes enregistrent les vaccins via le système d'information du Ministère de la Santé et de la Protection Sociale. L'objectif de ce système est d'effectuer le suivi de chaque per-sonne, afin de vérifier le

respect du schéma de vaccination dans le meilleur délai et de réduire ainsi les erreurs d'enregistrement pour améliorer la qualité de l'information.

Le système d'information est disponible pour les IPS sur le territoire national mais ne permet pas aux citoyens d'accéder à leurs vaccins numériquement ni de fournir des informations supplémentaires. De plus, la ligne directrice suggérée à l'utilisateur, à tra-vers son site web, est la suivante : (https://www.minsalud.gov.co/salud/publica/Va-cunacion/Paginas/pai.aspx) :
"N'oublie pas d'apporter la carte de vaccination. Si, pour une raison quelconque, vous ne l'avez pas, vous devriez vous rendre au point de vaccination le plus proche pour faire examiner votre cas d'une manière particulière "

Ces villes n'ont pas un outil ou une application mobile ni pour les citoyens ni pour les établissements de santé qui leur permit de gérer les communications avec les usagers.

2.3 QUELQUES EXEMPLES D'APPLICATIONS MOBILES DE VACCINATION

Après avoir passé en revue les expériences d'applications mobiles publiées sur des plates-formes telles que App Store et Play Store, nous pouvons renseigner les suivants :

- InfoVacunas (Chili)

Une application créée par le Ministère de la Santé du Chili qui permet d'enregistrer des informations relatives aux vaccins qui font partie du programme de vaccination, en plus de fournir des informations sur les points de vaccination les plus proches. Cette appli-cation vise à faciliter le processus de vaccination pour les parents et les professionnels en charge du programme national de vaccination du ministère de la Santé du Chili.

L'application est conçue pour fournir des informations au public sur les questions de vaccination et comprend un module pour la planification des vaccins à venir, qui est alimenté par chaque citoyen au moment du téléchargement de l'application et n'est con-necté à aucun système d'information nominale.

Ici le lien dans PlayStore: https://play.google.com/store/apps/details?id=cl.ceisu-fro.infovacunas

- Vacunas 3.0 (Espagne)

Cette application mobile fournit des informations liées à la vaccination des enfants et des adultes. On y trouve les recommandations de l'OMS et de l'Association espagnole de pédiatrie, ainsi que les effets secondaires possibles après l'application de produits biologiques. L'application est gratuite et disponible pour iOS et Android.

Comme le reste des expériences passées en revue, cette application ne dispose pas d'informations personnalisées et n'est pas non plus connectée aux bases de données de aucun prestataire de santé, car elle ne contient que des données générales.

Ici le lien dans PlayStore: https://play.google.com/store/apps/details?id=es.everywaretech.vacunas

3 LE PROJECT VACAPP

Les progrès de ce projet et le développement de cette application mobile visent à unir les forces pour atteindre les objectifs du Département de Santé du District (ou la ville de Bogotá) ainsi que à mettre en œuvre les directrices du Ministère de la Santé et de la Protection Sociale concernant la vaccination. Les objectifs sont énoncés comme suit :

1. Atteindre au moins 95% de couverture dans tous les produits biologiques faisant partie du régime de vaccination dans les populations cibles du programme.
2. Maintenir l'éradication de la poliomyélite dans la ville (Bogotá); consolider l'éli-mination de la rougeole, de la rubéole, du syndrome de la rubéole congénitale et du tétanos néonatal; contrôler l'incidence des cas de fièvre jaune, de diphtérie, de tuberculose méningée, d'hépatite A et B, de pneumocoque, d'Haemophilus influenzae type b, de diarrhée causée par le rotavirus, de coqueluche, d'oreillons, de grippe et de varicelle. En outre, réduire l'incidence du cancer du col de l'uté-rus grâce à la vaccination contre le virus du papillome humain.
3. Se conformer aux objectifs et stratégies définis dans le Plan stratégique pour l'éradication de la poliomyélite au niveau mondial, 2013-2018.
4. Vacciner la population sensible à la rougeole et à la rubéole dans le groupe d'âge de 2 à 10 ans.
5. Se conformer au Plan national de contrôle de l'hépatite virale, 2014-2017.

Cependant, atteindre la conformité avec ce qui précède, présente plusieurs inconvé-nients car :

- A ce moment-là, les utilisateurs ne savent pas en temps voulu les prochaines dates de vaccination selon leur schéma.
- La couverture vaccinale est faible.
- Il n'y a pas de communication bidirectionnelle entre le Département de la Santé de la ville de Bogotá et les utilisateurs vaccinés.
- Il n'y a pas assez d'informations sur les sites web et points de vaccination.
- Manque de connaissances de la part des utilisateurs sur les avantages de suivre un programme de vaccination.

Compte tenu de ce qui précède, le développement de VacApp devrait envisager un ensemble de spécifications techniques que, pour l'instant, nous pouvons énoncer de ma-nière non exhaustive à travers les objectifs suivants :

Objectif général : Créer une application mobile pour les citoyens qui permettra le suivi en temps réel des programmes de vaccination pour les enfants et les adultes dans la ville de Bogotá.

L'idée est de synchroniser le fonctionnement de VacApp avec le PAI du Secrétariat de Santé de la ville de Bogotá, de manière à permettre la visualisation des informations contenues dans ce système d'information, ainsi que l'envoi d'alertes, de journées de vac-cination et d'informations divers autours de la vaccination.

Objectifs spécifiques :

- Augmenter les informations sur l'environnement et le langage de programmation pour le développement de VacApp.
- Permettre à l'utilisateur de consulter son propre schéma de vaccination.
- Recevoir l'envoi des notifications en fonction des prochaines dates de vaccination.
- Montrer les informations des points de vaccination et inclure les heures d'attention, le téléphone et l'adresse.
- Créer un module qui permet d'envoyer des nouvelles et des informations d'intérêt pour les utilisateurs de l'application.

- Faire la publication de VacApp sur les serveurs du Département de Santé de la ville de Bogotá afin que les citoyens puissent le télécharger et l'installer sur leurs appa-reils mobiles.

Cependant, le développement de VacApp, en plus de faire face aux défis inhérents à toute application mobile, doit résoudre certains problèmes de gestion et de communi-cation entre les différentes entités de santé qui interviennent dans le processus de vac-cination:

- Sécurité dans le traitement des informations sensibles.
- Faiblesse dans le traitement de la loi Habeas Data
- Le manque de collaboration entre les entités nationales telles que le Ministère de la Santé et les Prestataires de Santé (EPS/IPS) pour l'appropriation, l'utilisation et la visi-bilité de l'outil technologique par les patients.

4 DEVELOPPEMENT D'VACAPP

Selon la recherche effectuée dans le Département de la Santé de la ville, nous avons observé que les développements PAI (ses différentes versions) ont été faits en utilisant des outils Microsoft .NET et que la base de données utilisée est un SQL Server 2014.

Pour le développement de VacApp, et après avoir effectué une étude des langages de programmation proposés par le marché, l'utilisation de l'environnement de développe-ment Xamarin de Microsoft a été déterminée, compte tenu des avantages suivants.

1. Le Secrétariat de Santé a déjà la licence requise pour la publication de l'appli-cation et le reste de ses développements sont dans les outils Microsoft.
2. Xamarin prend en charge la connexion à la base de données SQL Server qui gère les informations des personnes vaccinées et qui sera la principale source de données pour VacApp.
3. Il est possible de développer VacApp dans les trois principales plateformes du marché (iOS, Windows Phone et Android), en utilisant le même langage de pro-grammation, C #.
4. Avec Xamarin, nous obtenons 100% d'applications natives, comme si elles avaient été codées en Swift ou en Java, pour garantir la stabilité de l'application.
5. Couvrir l'ensemble du cycle de vie d'une application : le design, le développe-ment et les tests à la production, principalement.

Fig. 2. Infrastructure pour héberger VacApp dans le Secrétariat de Santé de Bogotá

La figure 2. montre un schéma initial d'infrastructure qui, pour héberger l'application dans le Secrétariat de Santé de Bogotá.

La base de données [6] est encadrée dans le modèle de qualité des données ISO 25012 (2009g ISO) et ISO 8000 (ISO 2014) en tant que cadre de référence pour définir et mesurer la qualité des données.

Norme internationale pour la qualité des données - ISO 8000 :

- Les principes de la qualité des données
- Les caractéristiques des données qui déterminent sa qualité
- Les protocoles qui assurent la qualité des données

Norme pour la définition et l'échange de données de base ISO – 22745

- Définir les dictionnaires techniques ouverts
- Application des dictionnaires aux données de base.

5 CONCLUSIONS ET TRAVAUX FUTURS

Actuellement, le projet est en étude de faisabilité économique. Cependant, du point de vue technique, le projet présente une avance majeure auprès de l'étape de design, ainsi comme les outils qui seront utilisés pour le développement.

Ce projet peux aussi donner lieu à l'utilisation d'outils de BigData, étant donnés la grande quantité de données qui seront gérés et manipulées. Donc, des outils de BigData seront nécessaire pour aider à la gestion et à la prise de décisions dans le domaine de la Santé en général, et de la vaccination en particulier.

REFÉRÉNCES

1. Ministerio de Salud y Protección Social. (2016) Lineamientos para la Gestión y Administración del Programa Ampliado de Inmunizaciones PAI. Recuperado de: https://www.minsalud.gov.co/si-tes/rid/Lists/BibliotecaDigital/RIDE/VS/PP/PAI/lineamientos-pai-2017.pdf
2. MINTIC. (2015). Manual Estrategia de Gobierno en Línea, 74. Recuperado de: http://estrategia.go-biernoenlinea.gov.co/623/articles-7941_manualGEL.pdf
3. Congreso Nacional de la República de Colombia. (2014). Ley 1712 de 2014, Por medio de la cual se crea la ley de transparencia y del derecho de acceso a la información pública nacional y se dictan otras disposiciones. Recuperado de: http://wsp.presidencia.gov.co/Normativa/Leyes/Documents/LEY 1712 DEL 06 DE MARZO DE 2014.pdf
4. Congreso de la República de Colombia. Ley estatutaria 1581 de 2012. Disposiciones Generales para la Protección de Datos Personales (2012). Recuperado de: http://www.secretariasenado.gov.co/se-nado/basedoc/ley_1581_2012.html
5. Congreso de la República de Colombia. (2012) Ley estatutaria 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. Disponible en: http://www.secretaria-senado.gov.co/senado/basedoc/ley_1581_2012.html
6. Jeffrey A. Hoffer, V. Ramesh, Heikki Topi. (2011) Modern DataBase Management, Pearson Educa-tions, Inc.
7. Carreño G, (2013). Fundamentos de Bases de Datos. Unidad Uno. Normalización. Colombia: Bogotá. Material didáctico Politécnico Grancolombiano.
8. Ministerio de Salud y Protección Social. (2018) Lineamientos para la gestión y administración del pro-grama ampliado de Inmunizaciones - PAI- 2018. Disponible en: https://www.minsalud.gov.co/si-tes/rid/Lists/BibliotecaDigital/RIDE/VS/PP/PAI/lineamientos-pai-2018.pdf
9. Alcaldía Mayor de Bogotá D. C. Plan Estratégico de Tecnologías de Información y Comunicación 2016
10. - 2020. Secretaría de Salud. Disponible en: http://www.saludcapital.gov.co/Planes%20Estrategi-cos/11_SDS_TIC_PL_002_Plan_Estrategico_Tecnologias_de_Informacion_y_Comunidades.pdf
11. ecretaría Distrital de Salud. Plan de Salud Pública de Intervenciones Colectivas 2017. Secretaría de Salud. Disponible en:http://www.saludcapi-tal.gov.co/Su_GPAISP/Caja_de_herramientas/1_ESTRATEGIAS_DE_INTERVENCION/3_ESTRA TEGIAS_INTERVENCI%C3%93N/PAI/PAI.pdf

PARALLEL AND DISTRIBUTED PROCESSING FOR UNSUPERVISED PATIENT PHENOTYPE REPRESENTATION

GARC´IA H. JOHN A.[1], PRECIOSO FRÉDÉRIC[1], STACCINI PASCAL[2], AND RIVEILL MICHEL[1]

1 UNIVERSIT´E CO^TE D'AZUR, CNRS, LABORATOIRE I3S, SOPHIA ANTIPOLIS, FRANCE

HENAO@I3S.UNICE.FR, {FREDERIC.PRECIOSO, MICHEL.RIVEILL}@UNICE.FR

2 UNIVERSIT´E CO^TE D'AZUR, CHU NICE, NICE

PASCAL.STACCINI@UNICE.FR

ABSTRACT.

The value of data-driven healthcare is the possibility to de-tect new patterns for inpatient care, treatment, prevention, and compre-hension of disease. Modeling precise patients phenotype representation from clinical data is challenging over its high-dimensionality, noisy and missing data to be processed into a new low-dimensionality space. Like-wise, processing unsupervised learning models into a growing clinical data raises many issues, in terms of algorithmic complexity, such as time to model convergence and memory capacity. This paper presents Diag-noseNET framework to automate patient phenotype extractions and ap-ply them to predict different medical targets. It provides three high-level features: A full workflow orchestration into stage pipelining for mining clinical data and using unsupervised feature representations to initial-ize supervised models; A data resource management for training parallel and distributed deep neural networks; and energy-monitoring tool for workload hardware characterization. As a case of study, we have used a clinical dataset from admission and hospital services to build a gen-eral purpose inpatient phenotype representation to be used in different medical targets, the first target is to classify the main purpose of in-patient care. The research focuses on managing the data according to its dimensions, the model complexity, the workers number selected and the memory capacity, for training unsupervised staked denoising autoen-coders over a Mini-Cluster Jetson TX2; therefore, mapping tasks that fit over computational resources is a key factor to minimize the number of epochs necessary to model converge, reducing the execution time and maximizing the energy efficiency.

Keywords: Health Care Decision-Making · Unsupervised Representa-tion Learning · Distributed Deep Neural Networks

1 INTRODUCTION

A critical step of personalized medicine is to develop accurate and fast artificial intelligence systems with lower rates of energy used for tailoring medical care (eg. treatment, therapy and usual doses) to the individual patient. In this context, in-ferring common patient phenotype patterns that could depict disease variations, disease classification and patient stratification, requires massive clinical dataset and computationally intensive models [1, 2]. Thus, the complex structure, noisy and missing data from large Electronic Health Records (EHR) data became a core computational task to automated phenotype extractions [3].

In this paper, we describe the unsupervised learning method for mining EHR data and build low-dimensional phenotype representations using a mini-cluster with 14 Jetson TX2 to distributed training and obtaining a patient phenotype representations used as input of supervised learning

algorithms for prediction the main purpose of inpatient care. We present an application-framework called DiagnoseNET that provides three high-level features: The first one enables a phenotypic discovering workflow orchestration into a stage pipelining, as mining EHR data, unsupervised representation learning and supervised learning; the second is a data resource management to feeding the clinical dataset into the Jetson TX2 according with their memory capacity, while multiple replicas of a model are used for minimizing the loss function and third, an energy-monitoring tool for scalability analyses impact of using different batch size factor to minimize the number of epochs needed to converge and projected the energy efficiency measures.

2 RELATED WORK

In the past century, health research models were traditionally designed to iden-tify patient patterns given a single target disease, where domain experts su-pervised definitions of the feature scales for that particular target and usually worked with small sample size, which were collected for research purpose [4, 5]. Nevertheless, in general clinical data are noisy, irregular and unlabeled for dis-covering directly the underlying phenotypes, this supposed a limitation for that approach. Nowadays, computer science has facilitated the design and implemen-tation of emerging frameworks and practical approaches, offering different ways to extract valuable information as phenotypes [6].

Derive patient phenotypes, it is necessary to extract the occurrence of their medical data (as, demographic, medical, procedures codes, etc) and their se-quential information in the time. A used method is *vector based representation* in which, for each medical target is constructed a matrix correlation between patients and groups of medical features [7], in which the sequence time is a limitation. A couple of are *nonnegative matrix factorization* and *nonnegative tensor factorization* for extracting phenotypes as a set of matrix, tensor candidates that shows patients clusters linked on specific medical features and their date [8–10]. Other approaches use non-negative vectors for embedding the clinical codes and use word representations as (skip-gram or Glove) to generate the corresponding visit representation [11].

Nevertheless, after success of unsupervised feature learning for training un-labeled data to dimensionality reduction and learn good general features repre-sentations and used either as input for a supervised learning algorithms [12], the application of employ it for produce patient phenotype representations can signif-icantly improve predictive clinical model for a diverse array of clinical conditions as it was shown in deep patient approach [13]. Other derivative approaches uses a record into a sequence of discrete elements separated by coded time, in which uses the unsupervised embedding Word2Vec to pre-detected the continuous vec-tor space, them uses a convolution operation which detects locals co-occurrence and pooled to form a global feature vector, which is passed into a classifier [14]. Another approach train a recurrent neural network with attention mechanism to embed patient visit vector to a visit representation, whic is then fed to a neural network model to make the final prediction [15].

However, the se approaches to the development of phenotyping algorithms demand considerable effort in deploying preprocessing pipelines and data trans-formation, in which are built without taking into account the response time. In this perspective, a great many commercial and academic authors have explored scaling up deep learning networks, training well-known datasets focused on the impact of synchronization protocol and state gradient updates [16–18]. At the same time, other groups have been working on high-level frameworks to easly scale out to multiple machines to extend libraries

for parameter management to allows more agile development, faster and fine tuning hyperparameter explo-ration [19,20]. All these developments are not applied to medical care and do not consider energy consumption. Our aim is to construct a completed framework to scaling deep learning techniques in direction of extracting effective patient phenotype representations on low-power platforms for empowering the hospitals and medical centers their ability to monitor health, to early disease detection and manage to personalise treatments to specific patient profiles.

3 MATERIAL AND METHODS

To illustrate phenotypic discovering workflow orchestration implemented in Di-agnoseNET application-framework. Which goal is to create a new lower dimen-sional patient phenotype representation for clinical predictive modelling, con-sider the DNN workflow orchestration into stage pipelining shown in Figure

1. The first stage *mining EHR data* to drive a binary patient term-document matrix from clinical document architecture. The second stage *unsupervised rep-resentation* for mapping the binary patient representation through unsupervised stacked denoising autoencoder to get a new latent space and identify patient phenotype representations. And the third stage *supervised learning* we use the latent patient representation as input for random forest classifier, and as initial izer for deep neural networks and the results are compared versus the binary patient representation

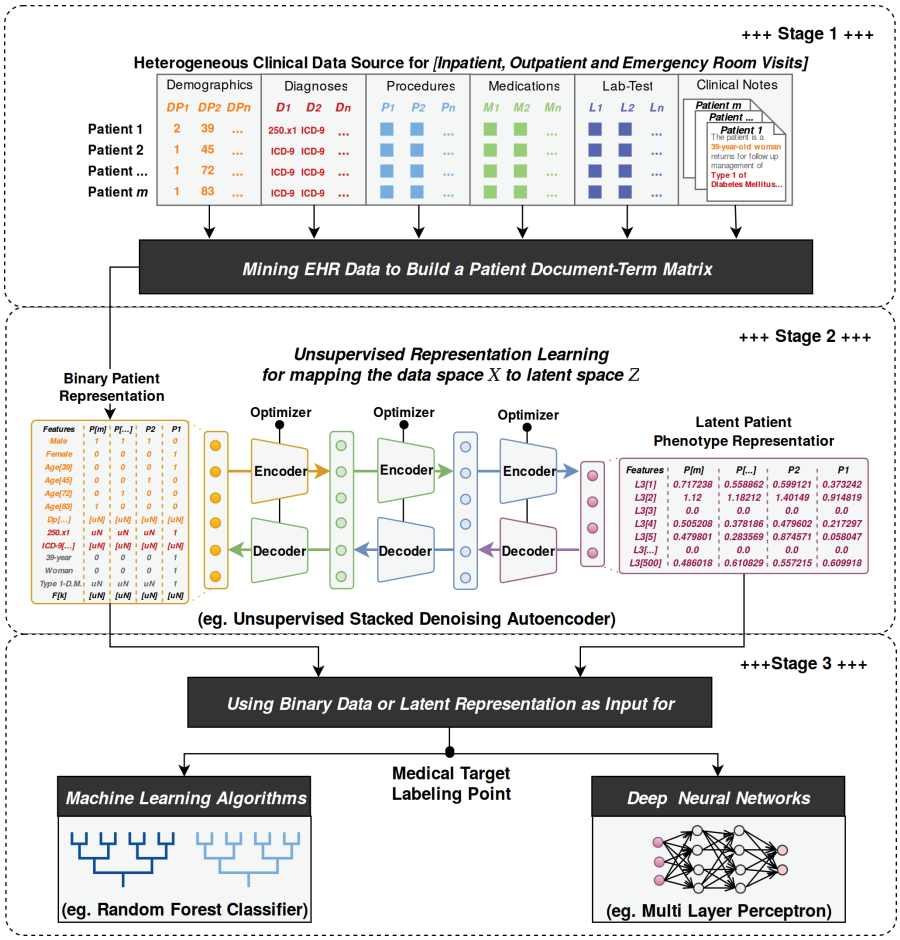


Fig. 1: Workflow scheme to automate patient phenotype extractions and apply them to predict different medical targets.

MINING EHR DATA

The growing health-wide research is largely due to clinical dataset are composed a secondary usage of patient records collected in admission and hospital process [21].

Therefore the EHR is not a direct reflection of patient and their physiology but is a reflection of recording process inherent in healthcare with noise and feedback loops [22].

A data mining library has been built as a collection of functions to feature extraction and to build a patient document-term matrix from a clinical dataset composed of discrete objects as diagnosis, procedures in ICD-10 codes, CCAM codes and other derived objects as admission hospital details represented in codes established by the agency ATIH and generated by the system PMSI for standardized the information contained in the patient's medical record.

The collection functions are:

1. Clinical Document Architecture (CDA): identifies the syntax for clinical records exchange between the system PMSI and DiagnoseNET, through the new versions generate by the agency ATIH. The cda schema basically con-sists of a header and body:
- Header: Includes patient information, author, creation date, document type, provider, etc.
- Body: Includes clinical details, demographic data, diagnosis, procedures, admission details, etc.
2. Features Composition: Serializes each patient record and get the CDA object for processing all patient attributes in a record object.
3. Vocabulary Composition: Enables dynamic or custom vocabulary for select-ing and crafting the right set of corresponding patient attributes by medical entities.
4. Label Composition: This function get the medical target selected from the CDA schema to build a one-hot or vector representation.
5. Binary Record: Mapping the features values from record object with the corresponding temrs by each feature vocabulary, to generate a binary corpus using Term-document Matrix.

UNSUPERVISED REPRESENTATION LEARNING

After the significant sucess of reesentation learning to encode audio, images and text with rich, high-dimensional datasets [**?**, 23–25].

In this work we extend the deep patient aproach [13], in which all the clinical descriptors are grouped in patient vectors and each patient can be described by a high-dimensional vector or by a sequence of vectors computed in a predefined temporal windows.

The collection of vectors are used to derive the latent patient representation through unsupervised encoder network, that pre-training autoencoder (the noiser) to get each hidden representation and transfer them to stacked encoder network and use clean data as input for mapping the previously trained weights respectively, as shown in the Figure 2.

This unsupervised encoder network is composed using *Unsupervised Stacked Denoising Autoencoders*: a deterministic mapping from cleaning partially cor-rupted input $\tilde{x}$ *(denoising)* to obtain a hidden features representation $y = f\theta(\tilde{x})$ by layer.

Therefore, each stacked layer is independently trained to reconstruct a clean input x from a corrupted version of it $z = g\theta^t(y)$, this approach was introduced by [26]. Previously each encoder was pre-trained to get a semantic representation parameters θ^t by denoising autoencoder which was trained before for obtaining a robust representation $y = f\theta(\tilde{x})$ from a corrupted input $\tilde{x}$. This is represented by the next steps:

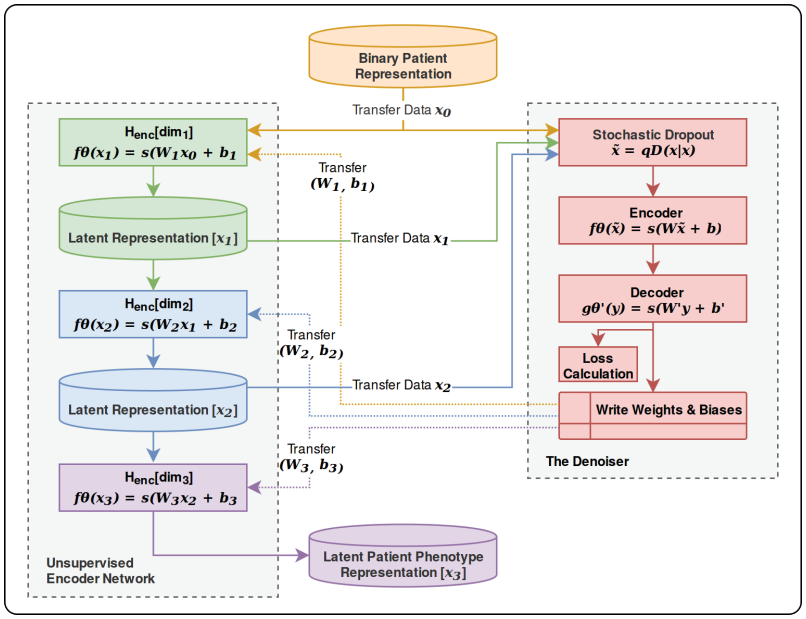


Fig. 2: scheme to illustrate the unsupervised encoder network for mapping binary pa-tient representation x to latent patient phenotype representation z.

1. Applied dropout to corrupting the initial input x into x˜ for the stochastic mapping x ∼ qD(x|x). The corrupted input is mapped as traditional autoencoders to get a hidden
2. representation y = fθ(x˜) = s(W x˜ + b).
3. Reconstruct a shematic representation of the procedure z = gθt(y) = s(W ty+ bt).
4. Where the parameters θ ∧ θt are trained to minimize the averange recontruc-tion error over training set, to have z as close as possible to the uncorrupte input x.
5. and this share the new semantic representation parameters θt to next layer as new initial input x2 and corrupting it into x˜2 by stochastic mapping x2 ∼ qD(x2|x2) and repeat steps.

SUPERVISED LEARNING

It is well known the performance of machine learning algorithms generally de-pends on data representations. for this reason, the resulting of deep features can be used either as input to a standard supervised machine learning predictor or as initialization for a deep supervised neural networks [12].

In this stage, we enrich the latent patient phenotype representation with the one-hot or multi-hot labels selected according to the applicative clinical target. then the implementations of random forest classifier and multlayer percetron network are used to com-pared the performance of these latent patient phenotype representation versus the same fine tuning hyperparameters of the classifier and the network using the binary patient representation.

PARALLEL AND DISTRIBUTED PROCESSING FOR TRANING DNN

The data resource management is being built as a high-level framework of ten-sorflow library, for scaling deep learning techniques in direction of optimally mapping the computational resources and adjust task granularity according with memory host capacity, model complexity and data batch size to minimize the energy consumption at training stage, as shown in the Figure 3.

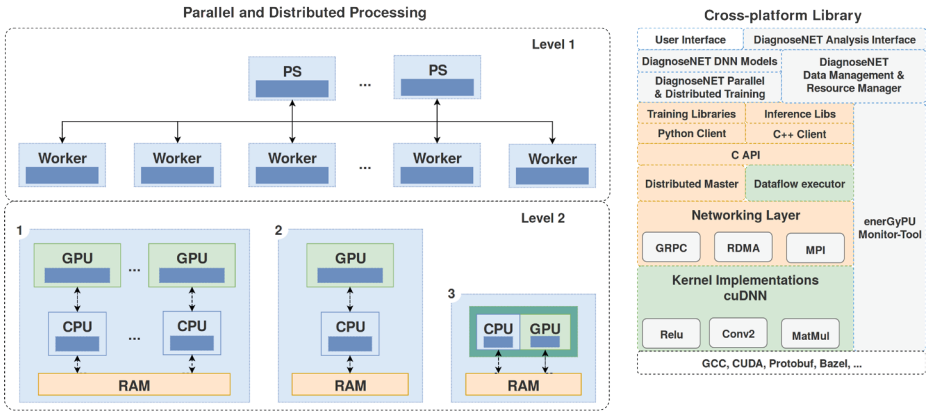


Fig. 3: Data resource management for training parallel and distributed deep neural networks and energy-monitoring tool.

The first approach uses data parallelism to training the unsupervised stacked denoising autoencoders on a mini-cluster Jetson TX2. Where all replicas task read the same values for the current DNN parameters, compute gradients in par-allel, and theirs apply together [27].

For training the To exploit the parallel and distributed processing To exploit the computational resources and the memory capacity of the Jetson TX2, the binary patients data is split acording with the number of Jetson boards and each part of the assigned binary batch data is again divided into the each host according with their memory capacity.

Concurrently the data resource manager designates the role and transmits the parameters to be used on the master (or parameter-server) and each of the workers.

DATASET

The clinical dataset from admission and hospital services have an average of 785.801 inpatients records by year, with records of hospitals activities in PACA and activities of residents, PACA hospitalised in another region. In which con-tains morbidity information, medical procedures, admission details and other variables, recorded retrospectively at the end of each week observed.

This in-formation may vary from one week to the next, depending on the evolution of the patient's clinical condition and management. As a case of study the clinical dataset taken 100.000 inpatients records by the year 2008 divided in 85.000 for training, 4.950 for validation and 10.050 for test with 11.466 features or clinical descriptors.

MEDICAL TARGET: CLASSIFICATION OF CARE INPATIENT PURPOSE

The first medical target is to classify the main purpose of inpatient care repre-sented as ICD-10 codes. The PMSI system can be assigned ICD-10 codes of the *Care Inpatient Purpose* as a high-level entry called Clinical Major Category used for billing procedures. The first exploration is used the Clinical Major Category to evaluate the classification (e.g 23 labels by the year 2008). An example of the first medical labels, consider the following inpatient cases shown in 2.

Table 1: Hierarchization of diagnosis-related group to selected the clinical major cate-gory as labels linked with the care inpatient purpose.

	Diagnosis-related Group	ICD-10 Codes	Definition
Patient 1	Morbidity Principal	R402	Unspecified coma
	Etiology	I619	Nontraumatic intracerebral hemorrhage, unspecified
Medical Target	Care Purpose	Z515	Encounter for palliative care
Label used	Clinical Major Category	20	Palliative care
Patient 2	Morbidity Principal	R530	Neoplastic (malignant) relate fatigue
	Etiology	C20	Malignant neoplasm of rectum
Medical Target	Care Purpose	Z518	Encounter for other specified aftercare
Label used	Clinical Major Category	60	Other disorders

4 EXPERIMENTS AND RESULTS

The first and second sessions of experimentation studies an established-set of common hyperparameters with different batch sizes to examine the relationship between a network's convergence time, energy consumption and its reliance to generate a low-dimensional space as a latent patient phenotype representation. The third session analyzes the scalability to execute one network model with dif-ferent batch sizes for DNN distributed training on a different number of workers (Jetson TX2 boards), in order to look a workload hardware characterization.

1. Number of gradient updates as factor to early model convergence.
2. Model dimensionality as factor to generate quality latent representation.
3. Number of workers and task granularity as factor to early model convergence on synchronous distributed processing.

Number of gradient updates as factor to early model convergence

The relationship between a mini batch size and network's convergence time is linked with the maximun number of state gradient updates by epoch. Thus, it is projected into execution time and energy consumption limited for the mem-ory capacity and data transfer (from host memory to device memory), when using GPUs as an accelerator to train the model.

The experiment uses the traditional fully conected autoencoders, parameterized with 3-hidden layers of [2048, 2048, 500] neurons per layer, the Relu is used as activation function; Adam such as optimizer, sigmoid cross entropy as loss function. The clinical datased uses 84.999 records for training and 4.950 records for validation, the next Figure presents the results of validation dataset.

Each mini bacth partition factor process an epoch of the network on aver-age of [21.73, 23.82, 26.61] $seconds$ respectively, showing a slight increase when loading a larger number of mini batches.

However, the larger mini-batch parti-tion requires processing a greater amount of epochs to locate the convergence point as 100, 20 and 10 epochs respectively. Thus, the projection for training the network presents an average power consumption of [63.35, 86.61, 82.21] $watts$ respectively with an energy consumption of [137.65, 41.26, 21.87] $Kilojoules$, be-ing 768 the most energy efficient factor to generate the batch gradient updates, as shown in the Figure **??**.

In which is observed that using large number of records per batch generates idle status on GPU Streaming Multiprocessor (SM) with a SM frequency of 847.49 MHz, while the data is transferred from the host memory to device memory, unlike the other two got a 1071.97 and 1015.49 MHz respectively.

MODEL DIMENSIONALITY AS FACTOR TO GENERATE QUALITY LATENT REPRESENTATION

The analyzes studies the relationship between a model complexity, network's converge time and its reliance to generate a low-dimensional space as a latent patient phenotype representation.

Specifically the experiment comparison us-ing a established-set of hyperparameters on 3 model variations for each net-work to compare the sigmoid vs relu as activation function to generate the latent representation, using three variations of number of neurons per layer as [4086, 2048, 768], [2000, 1000, 500] and [500, 500, 500].

The evaluation of accu-racy is measured comparing the latent representation generated by each network model and used as input for random forest classification of the clinical major category on 23-labels and their energy efficiency.

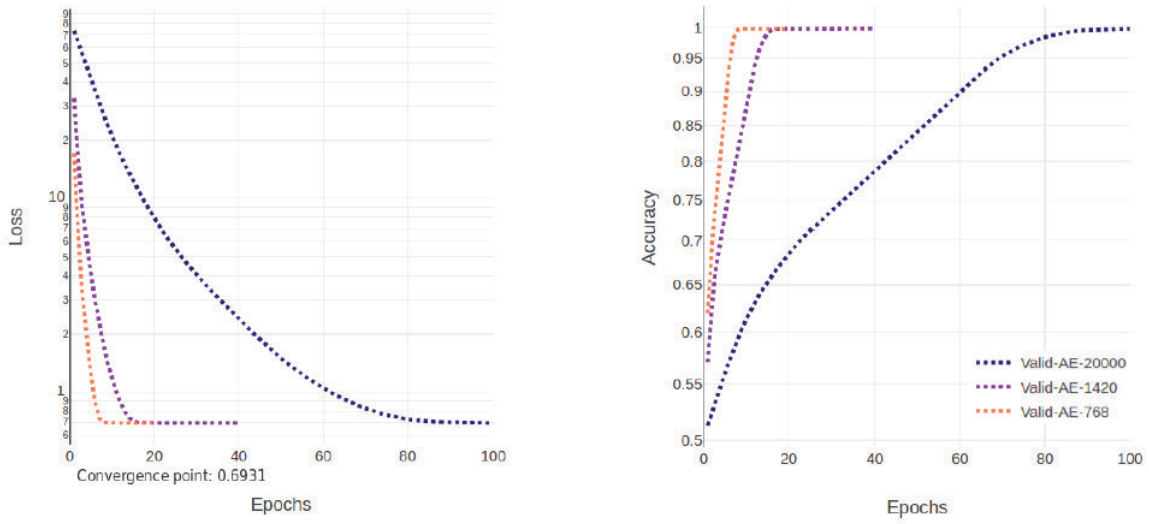
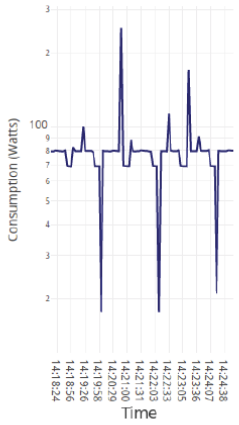
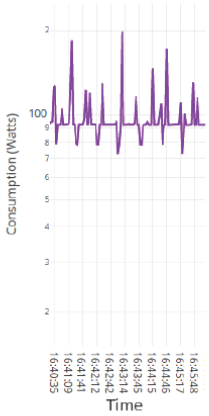


Fig. 4: Network convergence using batch partitions of [20000, 1420, 768] records to gen-erate [4, 59, 110] gradient updates by epoch respectively.

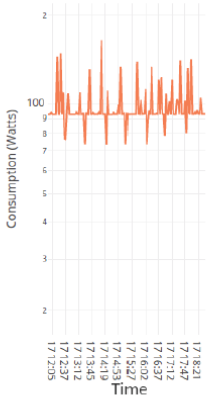
Fig. 6: Power consumption in a window of 6 minutes.



63.35 ¡*Watts* in average to process 68 gradient updtaes in 17 epochs.



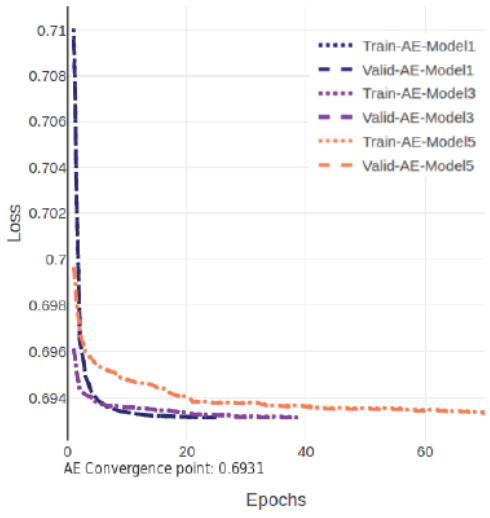
86.61 *Watts* in average to process 885 gradient updtaes in 15 epochs.



82.21 *Watts* in average to process 1540 gradient updtaes in 14 epochs.

1. The first network selected was a traditional fully conected autoencoders with 3 hidden layers to generate the latent representation.
2. The second is an End to End network using 3 hidden fully conected autoen-coders to generate the latent representation as input for the next 4 hidden multilayer percetron.
3. Encoder network using 3 hidden unsupervised stacked denoising autoen-coders to initialize the next 3 hidden layers to encode and generate the latent representation.

Fig. 7: Comparinson of different model dimensionality using sigmoid as function to generate the latent representation.



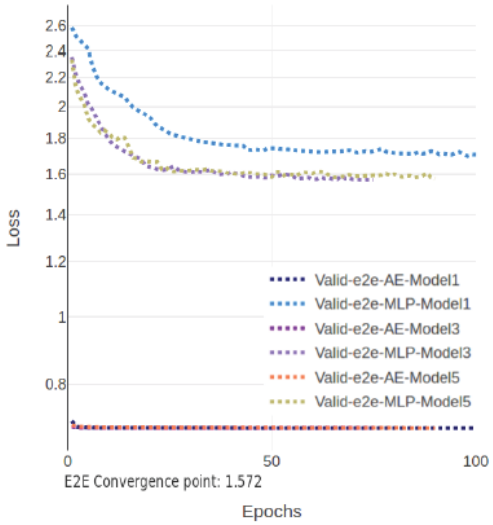


Fig. 9: Comparinson of different model dimensionality using relu as function to generate the latent representation

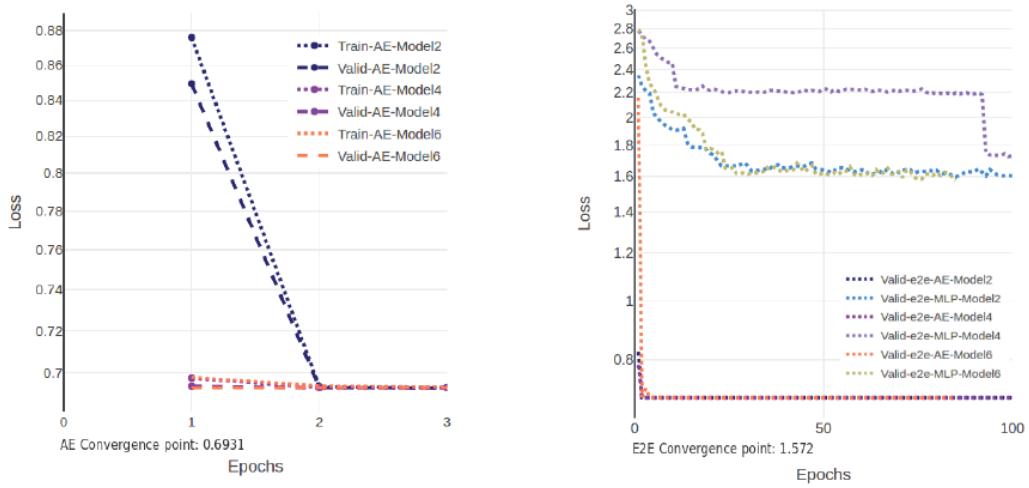
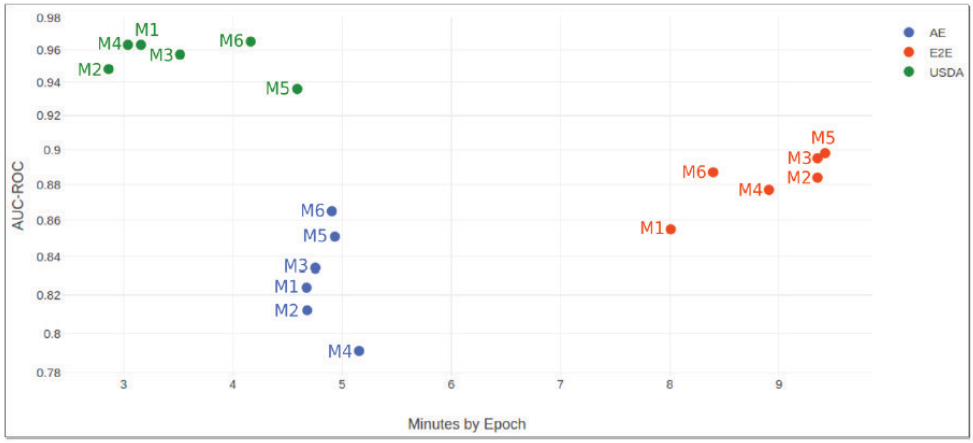


Fig. 10: Evaluation to classify the 23 labels using the latent representation as input for training random forest classifier and their execution time by epoch.



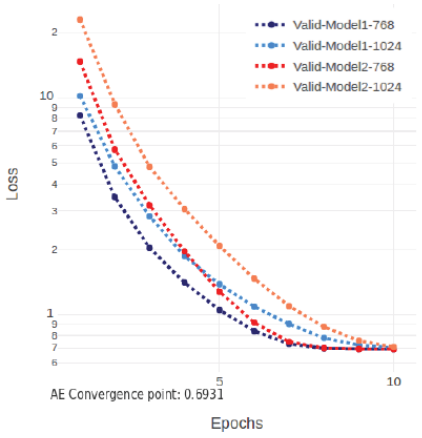
NUMBER OF WORKERS AND TASK GRANULARITY AS FACTOR TO EARLY MODEL CONVERGENCE

The experiment analyzes the scalability for training a traditional autoencoders using different number of workers using an established-set of hyperparame ters in two variations of number of neurons per layer as [2000, 1000, 500] and [2048, 1024, 768]. The different number of Jetson-TX2 Groups are:

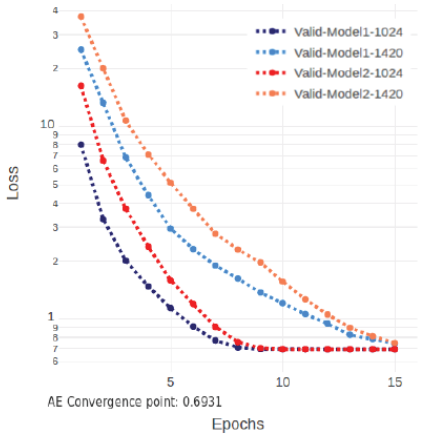
1. 1 P. Server and 3 workers – > Batch size: [768, 1024]
2. 1 P. Server and 6 workers – > Batch size: [1024, 1420]
3. 1 P. Server and 8 workers – > Batch size: [1066]

In this case is showed 1 PS and 8 workers processing in data parallelism and training the unsupervised encoder network for mapping binary patient rep-resentation x to latent patient phenotype representation z. Where shows the synchronous cooperation for going to converge point in the Figure.

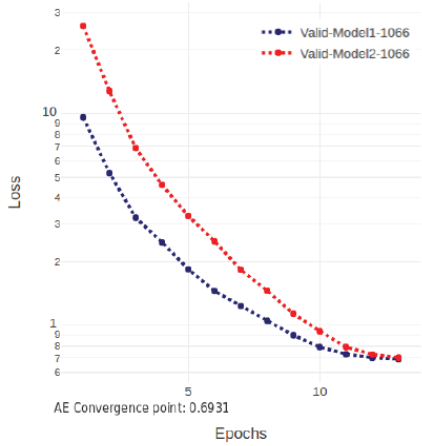
Fig. 11: Number of workers and task granularity as factor to early model convergence.



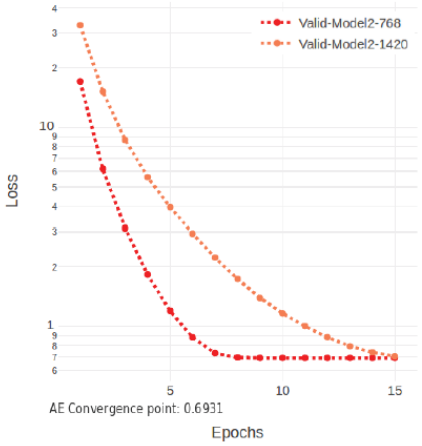
1.30 mins in avergange for processing one epoch on 1 PS 3 workers.



1 min in avergange for processing one epoch on 1 PS 6 workers.



50.6 secs in avergange for processing one epoch on 1 PS 8 workers.



25.75 secs in avergange for processing one epoch on 1CPU and 1GPU Titan X.

Table 2: Preliminary results for processing the unsupervised patient phenotype repre-sentation on the mini-cluster Jetson TX2.

AE	**1 PS and 3 Workers**		**1 PS and 6 Workers**		**1 PS and 8 Workers**		**1 CPU and 1 GPU**	
Network	Batch Fc.	Converge Time	Batch Fc.	Converge Time	Batch Fc.	Converge Time	Batch Fc.	Converge Time
M1	768	13.49 *mins*	**1024**	**9.95** *mins*	1066	10.18 *mins*		
M1	1024	11.90 *mins*	1420	10.51 *mins*				
M2	768	14.50 *mins*	1024	11.40 *mins*	1066	11.76 *mins*	768	3.97 *mins*
M2	1024	12.50 *mins*	1420	12.48 *mins*			1420	5.96 *mins*

CONCLUSIONS

The summary results to predict the first medical target can shows, that using the new lower-dimensional patient representation, reduces the number of sparse features to classify at stage 3.

In which, the execution time for training is min-imized by 41% with regard to binary patient representation and the precision to classify the first medical target and reduce significantly the task-complexity for training deep networks a supervised stage.

The unsupervised encoder net-work keeps the accuracy and reduce the classification time. To minimize the execution time for training DNN on Jetson TX Cluster, depends of the appli-cation capacity for mapping the data on computational resources data resource management into mini-batches. and the model complexity and gradient gener-alization is the important factor to minimize the number of epochs needed to model convergence.

ACKNOWLEDGMENTS

This work is partly funded by the French government labelled PIA program under its IDEX UCAJEDI project (ANR–15–IDEX–0001). The PhD thesis of John Anderson Garc´ıa Henao is funded by the French government labelled PIA program under its LABEX UCN@Sophia project (ANR–11–LABX–0031–01).

REFERENCES

1. Kathrin Heinzmann, Lukas Carter, Jason S. Lewis, and Eric O. Aboagye. Multi-plexed imaging for diagnosis and therapy. 1, 09 2017.
2. Cheng Yu, Wang Fei, Zhang Ping and Hu Jianying. Risk Prediction with Electronic Health Records: A Deep Learning Approach, 2016.
3. Lasko TA, Denny JC and Levy MA. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data, 2013.
4. Development of Inpatient Risk Stratification Models of Acute Kidney Injury for Use in Electronic Health Records. Medical Decision Making, 30(6):639–650, 2010. PMID: 20354229.
5. Kennedy E.H., Wiitala W.L., Hayward R.A., and Sussman J.B. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Medical Care, 2013.
6. Sheng Yu, Katherine P Liao, Stanley Y Shaw, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S. Kohane, and Tianxi Cai. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5):993–1000, 2015.
7. Xiang Wang, Fei Wang, and Jianying Hu. A Multi-task Learning Framework for Joint Disease Risk Prediction and Comorbidity Discovery. In *Proceedings of the 2014 22Nd International Conference on Pattern Recognition*, ICPR '14, pages 220–225, Washington, DC, USA, 2014. IEEE Computer Society.
8. Joyce C. Ho, Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics*, 52:199–211, 2014. Special Section: Methods in Clinical Research Informatics.
9. Ioakeim Perros, Evangelos E. Papalexakis, Fei Wang, Richard W. Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. SPARTan: Scalable PARAFAC2 for Large & Sparse Data. In *Proceedings of the 23rd ACM SIGKDD International Con-ference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 375–384, 2017.
10. Ioakeim Perros, Evangelos E. Papalexakis, Haesun Park, Richard W. Vuduc, Xi-aowei Yan, Christopher deFilippi, Walter F. Stewart, and Jimeng Sun. SUSTain: Scalable Unsupervised Scoring for Tensors and its Application to Phenotyping. *CoRR*, abs/1803.05473, 2018.
11. Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, and Jimeng Sun. Multi-layer Representation Learning for Medical Concepts. *CoRR*, abs/1602.05568, 2016.

12. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives, April 2014.
13. Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *SCIENTIFIC REPORTS*, 2016.
14. P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. *mathttDeepr*: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30, 2017.
15. Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Ji-meng Sun. GRAM: Graph-based Attention Model for Healthcare Representation Learning. *CoRR*, abs/1611.07012, 2016.
16. Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc'Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, and Andrew Y. Ng. Large Scale Distributed Deep Networks. In *NIPS*, 2012.
17. Janis Keuper and Franz-Josef Preundt. Distributed Training of Deep Neural Net-works: Theoretical and Practical Limits of Parallel Scalability. In *Proceedings of the Workshop on Machine Learning in High Performance Computing Environments*, MLHPC '16, pages 19–26, Piscataway, NJ, USA, 2016. IEEE Press.
18. Wei Zhang Fei Wang Suyog Gupta. Model Accuracy and Runtime Tradeoff in Distributed Deep Learning: A Systematic Study. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4854–4858, 2017.
19. Li Zhang and Yufei Ren Wei Zhang" "Yandong Wang. Nexus: Bringing Efficient and Scalable Training to Deep Learning Frameworks. In *25th IEEE International*