

Using Convolutional Neural Networks to Develop Local Exposure Models for Seismic Risk Assessment from Street Imagery

Daniela Gonzalez, Raúl Ramos-Pollán, Ana B. Acevedo, Juan C. Duque, Alejandro Betancourt, Sebastian García and Diego Rueda-Plata

Abstract—This work uses Convolutional Neural Networks (CNN) on Google Street View imagery to predict the structural typology of buildings to be used in local exposure models for seismic risk assessments. Exposure models are usually challenging to develop, since information is scarce and requires expensive and time consuming manual surveys. The method hereby proposed aims at lowering cost factors when developing exposure models and, thus, enabling their application to large urban settlements. We built a dataset of nearly 10,000 images extracted from Google Street View on the city of Medellín (Colombia), which were manually annotated by experts with their structural typology. We then created a predictive model for the two major structural typologies for buildings (reinforced concrete and masonry) by finetuning a state of the art publicly available CNN, and obtained an overall 85% in prediction accuracy, which is analogous to human expert's performance. We also explored in detail the model's performance in terms of geographical location and building subtypology in order to understand how it can be integrated within a seismic risk assessment process.

Index Terms—Seismic risk assessment, Convolutional Neural Networks, exposure model, street-view data, SDG 11.

I. INTRODUCTION

A n exposure model is a detailed description of the exposed assets in a region: properties, infrastructure, population and economic activities [1]; and it is a key input for seismic risk assessment [2]. Exposure models, along with seismic hazard and vulnerability models, are used to estimate the probability of losses if an earthquake takes place [3], [4].

Manuscript received XXX XX, 2018; revised XXX XX, 2018. Date of publication XXX XX, 2018; date of current version XXX XXX, 2018

This article was completed with support from the PEAK Urban programme, supported by UKRI's Global Challenge Research Fund, Grant Ref: ES/P011055/1. The authors also thank the support of the Apolo supercomputer resources and team at EAFIT and the financial support for database collection was given by EAFIT University, Grant Ref: 690-000026. The usual disclaimer applies

D. Gonzalez and A. B. Acevedo are with the Department of Civil Engineering, Universidad EAFIT, Medellín, Colombia (email: dgonza26@eafit.edu.co; aceved14@eafit.edu.co).

R. Ramos-Pollán is with the Universidad de Antioquia, Medellín, Colombia (email: raul.ramos@udea.edu.co)

J. C. Duque is with the Department of Mathematical Sciences (RiSE-group), Universidad EAFIT, Medellín, Colombia (email: jduquec1@eafit.edu.co).

A. Betancourt is with the Digital Department - Business Analytics, Ecopetrol, Bogotá, Colombia (email: alejandro.betancourt@ecopetrol.com.co).

S. García is with the Center of Excellence and Appropriation on Big Data and Data Analytics (CAOBA), Universidad EAFIT, Medellín, Colombia (email: sgarcia18@eafit.edu.co).

D. Rueda-Plata is with the Universidad Industrial de Santander, Bucaramanga, Colombia (email: ing.diegorueda@gmail.com).

The development of an exposure model is a challenging task, as many of the required information often does not exist or is not publicly available. When small settlements are considered, information may be obtained from in-situ surveys with a moderate use of human and economic resources. But, as the size of the settlements increases, so does the cost and time required to carry out these surveys [5], [6], [7], [8].

This paper uses Convolutional Neural Networks (CNNs) over Google Street View imagery for developing a residential building exposure model for the city of Medellín (Colombia), which is the second largest city of Colombia with a population of 2.5 million inhabitants. In particular, we evaluate the usage of CNNs to predict the structural typology of buildings, which is an indicator of building behavior when exposed to seismic loads.

We gathered and manually annotated a dataset of around 10,000 Google Street View images with buildings within the urban area of Medellín. Then, we used this data to fine-tune state of the art publicly available pre-trained CNNs and obtained 85% accuracy when predicting the first level class distributions of building typologies. This is quite close to human level accuracy and we believe it constitutes a solid foundation to further develop low-cost computer-assisted local exposure models.

In the context of Latin American cities, the most urbanized region in the world, the possibility of making seismic risk assessments in a shorter time and at a lower cost, is an important contribution to the achievement of sustainable development goal (SDG) 11, “Make cities and human settlements inclusive, safe, resilient and sustainable.” In this sense, Latin American cities must face the challenge of revisiting their built environment and increase its quality to make cities safer. In addition, exposure models are a key component of a seismic risk assessment, which is essential for understanding disaster risk, one of the four priorities of action of the Sendai Framework for Disaster Risk Reduction¹.

The rest of this paper is organized as follows. Section II reviews previous work done in this area. Section III describes the data collection process and the construction of the dataset. Section IV details the experimental section. The results and subsequent discussion are presented in section V. Finally, Section VI presents the main conclusions.

¹www.unisdr.org/we/coordinate/sendai-framework

II. PREVIOUS WORK

A. Buildings exposure models

Building exposure models include information about the geographical location of each asset (building) and its structural typology. The structural typology indicates the building behavior when exposed to seismic loads. For seismic risk assessment building typology is a function of the lateral load-resisting system and its materials, building height, date of construction, shape of the building plan, among others. The number of building typologies must be enough to represent the building stock of the location under study, keeping in mind that a fragility/vulnerability model —relationship between damage/cost and a strong motion measure— for each typology must be available [9].

Data collection is one of the main tasks on the development of an exposure model [10], and it is also one of the main challenges when implementing these models in developing countries, in which the lack of systematized information is a well known problem. For the city of Medellín, previous works have explored different sources of information including census data, cadastral information, in-situ surveys, and, recently, remote sensing.

Reference [2] uses census data as input for the development of an exposure model for the residential building stock of South America. The authors used for Colombia the census of 2005, which gives information on the predominant material of the floor and exterior walls. However, expert opinions were necessary to gather information, not available in the census, such as the lateral load resisting system. Cadastral information has been used in Medellín as a source for building plans and building heights [11], [12]. Both authors used additional information collected by in-situ surveys in order to define building typology, which is mainly defined by the lateral load resisting system. As the best possible exposure model would lead to the reduction of the uncertainty level on the risk assessment [8], the building classification needs to be performed by experts.

Since the early 2000s, the development of remote sensing technologies, and their high degree of penetration in urban environments worldwide, created an opportunity for gathering data for exposure models from the air [5], [13], [14], [15]. This technology allows for the measure of variables such as plan-built areas, building height, type of roof, building classification (in terms of in-plan area, number of stories, detached/attached house, etc.), and building age (by comparing images taken at different years). More recently, some authors have started to use Google Street View to replace fieldwork with virtual tours to visualize the facades and gather their lateral load resisting system and its material from the comfort of their desks [10], [16]. Reference [8] used Google Street View images to automatically retrieve and map floor numbers in urban buildings.

In this paper, we aim to go one step further by taking advantage of the developments in artificial intelligence, specifically within the area of Neural Networks, to develop a system capable to learn structural building typology from experts and

classify, automatically, a great amount of facades, captured from Google Street View.

B. Convolutional Neural Networks (CNN)

Deep learning methods have had in recent years a wide degree of success in a diversity of computer perception tasks [17]. They allow avoiding the design of specific feature detectors for images by looking for a set of transformations directly from the data. This approach has had remarkable results, particularly in computer vision problems such as natural scene classification and object detection [18].

Convolutional neural networks [19] are specialized architectures of artificial neural network targeted to process signals in general and, specifically, two-dimensional images. In a CNN, images pass through successive layers conforming a hierarchical pattern recognition system that, at the same time, learns what features are best suited to classify an image, recognize objects, etc.

The typical architecture of a CNN is a succession of convolutional layers accompanied by other supporting layers (such as pooling, dropout, etc.). The convolutional layers learn a set of filters or kernels that they activate when a pattern of features is present somewhere in the input image. Their convolutional and hierarchical nature provides certain scale and location invariance capabilities. Pooling layers progressively reduce the spatial size of the representation, diminishing the amount of parameters and computation in the network. Finally, a CNN is completed with fully-connected (FC) layers as a regular multilayer perceptron providing the network's output (class probabilities, etc.).

Besides simulation based approaches (see [6], [20]) and classical statistical or modeling methods ([21], [22], [23]), there has been seldom usage of machine learning methods in general for seismic risk assessment depending always on data availability. Certain machine learning approaches have been based on applying standard algorithms to data acquired from sensors or databases (public registries, etc.), such as the works in [24], [25] or [26].

On the other side, image processing approaches have been proposed in related areas, such as for earthquake damage estimation [27], exposure estimation [28], [29], landslide risk assessment [30], etc. But the vast majority is based on satellite imagery.

CNNs are starting to be used very timidly in the area. For instance, in [31] a CNN is used to process ground velocity records to detect earthquakes. However, CNNs have been used to extract information and classify Google Street View images. This includes, street number recognition [32], traffic sign recognition [33], building number of stories [8], or to estimate the demographic composition of neighborhoods [34].

III. DATA DESCRIPTION

A. City of Application: Medellín, Colombia

Between the early 1950s and the early 1970 Medellín's population grew from 358,189 to 1,077,252 inhabitants, mainly due to the forced displacement from rural areas to urban informal settlements on the valley slopes. After the urban



Fig. 1. Location of Medellín.

war against the drug cartels, from middle 1980s to the early 1990s, Medellín's annual population growth reached 2.6%. Today, Medellín is the second most populated city of Colombia with an estimated population of 2.5 million, in an area of 1,152 km², divided in 271 neighborhoods grouped in 16 urban districts (comunas).

Medellín (Fig. 1) is located in an intermountain valley at 1,460 m above mean sea level, in a medium seismic hazard zone [35]. The possibility of earthquake occurrence and the city size has made its seismic risk a concern on the last decades [2], [11], [12], [36], [37]. Construction quality of Medellín's building portfolio is closely related to inhabitant's acquisition level and building age: best construction practice is found in medium-high and high income zones. In addition, seismic design in Colombia has only been mandatory from 1984; therefore, few buildings built before that date can sustain seismic loads. On the other hand, a great percentage of Medellín building stock is informal construction in low-income zones, which do not fulfill code requirements [12] and are expected to have a poor performance under seismic loads.

B. Building typologies

An exposure model for seismic risk assessment involves the identification of different buildings typologies that have a fragility/vulnerability function associated with them. Each typology classifies buildings according to a set of characteristics such as material of the load resisting system, lateral load resisting system, building height, etc. [38], [39], [40]. In this paper we follow the building taxonomy for earthquake assessment developed by the Global Earthquake Model (GEM) Foundation [41]. The GEM taxonomy includes 13 attributes that cover four dimensions: Structural System, Building Information, Exterior Attributes, and Roof/Floor/Foundation. Table

I shows different options to gather each attribute. Direct observation (by experts) is the technique that allows for the measurement of most attributes, and it is the only technique for measuring lateral load-resisting system and material of the lateral load-resisting system. Lateral load-resisting system refers to the horizontal and vertical elements that transfer lateral seismic forces to the building foundations; it can be a system of walls, beams/columns, beams/columns/walls, etc. The lateral load-resisting system can be made of different materials such as reinforced concrete, masonry, steel, earth, etc.

In big locations like Medellín, it is not possible to survey each one of the assets; therefore, several assumptions need to be made for the implementation of an exposure model: some authors gather the building attributes from census data [2]; other authors take statistics from a sample of surveyed structures and used them to allocate the load resisting system of the non-surveyed buildings [11], [12], [36].

TABLE I
GEM BUILDING TAXONOMY AND METHODS FOR DATA COLLECTION

Attribute	Data collection method				
	Cadastral map	Remote sensing*	Direct observation	Census data	Other
Direction	✓	✓	✓		
Material of the lateral load-resisting system			✓		
Lateral load-resisting system			✓		
Height	✓	✓	✓		
Date of construction					(1)
Building position within a block	✓	✓	✓		
Shape of building plan	✓	✓	✓		
Occupancy			✓	✓	
Structural irregularity	✓	✓	✓		
Exterior walls			✓	✓	
Floor			✓	✓	
Roof		✓	✓		
Foundation					(2)

*: Remote sensing refers to the use of aerial photography.

(1): Comparison of multitemporal satellite images, building blueprints.

(2): Building blueprints.

Lateral load-resisting system of the building stock varies from location to location due to differences in construction practices, material availability, age of the buildings, weather, etc. For the city of Medellín, the high levels of socio-economic inequality are also reflected in a high urban heterogeneity. According to [11], “[Medellín] has a portfolio of buildings mainly comprised by low and intermediate rise masonry structures and reinforced concrete structures of medium and high rise in the developing residential areas.” Reference [16] identified nine building typologies using the GEM taxonomy based on three attributes: material of the lateral load-resisting system, type of lateral load-resisting system and ductility level. Building typology “others” was included in order to consider buildings built with unconventional materials; it constitutes an insignificant percentage of Medellín's building stock and it is excluded of the analysis in the present work.

The lateral load-resisting system of the eight considered typologies is either wall (LWAL), infilled frame (LFINF) or



Fig. 2. Building typologies for the residential building stock of Medellín (names according to GEM taxonomy). Images sources: photos taken by the authors in the city of Medellín. Jan., 2018.

dual frame-wall system (LDUAL). The system can be made of masonry or reinforced concrete (CR); in masonry buildings the lateral load-resisting system is composed of masonry walls (vertical planar elements), which can be unreinforced (MUR: masonry without any form of reinforcement), confined (MCF: construction in which masonry walls are first laid and the horizontal and vertical reinforced confining elements are cast), and reinforced (MR: masonry wall construction in which reinforcement is embedded in such a manner that two materials act together in resisting forces). Lateral load-resisting system of the RC buildings can be either frames, walls or a combination of both. A moment-resisting frame is a beams and columns structure with strong and rigid beam-to-column connections; when some bays of frames are infilled with masonry walls the system is known as infilled frame. In a wall system walls resist both gravity and horizontal forces; RC walls are monolithic, which is not the case of masonry walls. In a dual RC frame-wall system both frame and RC walls resist lateral loads. The third attribute for building classification, ductility, refers to the building capacity to sustain deformation without failure. Building typologies are classified as ductile (DUC) and non-ductile (DNO). Fig. 2 presents examples of each building typology.

According to [11], in Medellín, unreinforced masonry dwellings are common in low-income areas, and reinforced masonry and reinforced concrete dwellings are found in middle to high-income level areas. Although unreinforced masonry (MUR) is a building typology that is forbidden by the current seismic code [42] in medium seismic hazard zones like Medellín, more than half of the city building stock is MUR structures [12].

The height of buildings is usually related to its typology. Masonry buildings have a limited number of stories (commonly no more than five stories) due to material characteristics and code specifications. In the case of RC buildings, RC frames are commonly used up to ten stories. Due to code regulations, a higher number of stories require the use of walls; therefore, buildings with ten or more stories are either RC ductile wall systems or ductile dual RC frame-wall systems. RC frames can be ductile or non-ductile. Non-ductile RC frames are those buildings built without code regulations, either because they are informal constructions or they were built before 1984, year in which the first seismic code became

mandatory in Colombia [43]. Before 1984 it was not common the use of elevators, thus it can be assumed that the number of building stories for non-ductile RC frames is limited to approximately 6 stories.

In the broad sense, building typologies in Medellín can be grouped in two macro-classes: masonry buildings and RC buildings. Masonry buildings are limited to five stories, whereas RC buildings can have any number of stories. Classification efforts should concentrate in buildings up to five stories as they can be either masonry or RC. Although the macro classes by themselves do not indicate the exact lateral load-resisting system, they give valuable information of the material of the lateral load-resisting system. This information combined with additional information such as statistics from in-situ surveys, information of the year of construction, knowledge of the city development, etc. can be used for the final exposure model definition.

C. Buildings Dataset

This paper relies on the data set built by [16], which developed an exposure model for Medellín based on cadastral information, expert judgment and survey data. To construct the data set, the authors defined a set of non-spatially contiguous homogeneous zones by aggregating neighborhoods sharing similar characteristics in terms of socio-economic level and height of the buildings. Fig. 3 shows the spatial distribution of the homogeneous zones. Although the process of defining the zones does not include a spatial contiguity constraint, the resulting zones show clear spatial patterns that tell about the high level of social-urban segregation in the city.

Each zone has a code with three digits: the first digit is a number that indicates the predominant building height (1: 1-2 stories; 2: 3-4 stories; 3: 5-6 stories). The second digit is a letter that indicates the level of socio-economic strata used by the Government to charge the Utility fees (A: strata 1 and 2 or low income; B: strata 3 and 4 or middle class; C: strata 5 and 6 or high income; D: non-predominant strata). Finally, the third digit indicates the remaining building height (1: more than 80% of the buildings have the same height as the predominant height; 2: 1-6 stories).

Once the zones were defined, [16] used Google Street View to sample 11,381 buildings distributed throughout the entire city. An image of the facade of each building was collected using Google Street View. Building information along with the images was stored by the use of the Inventory Data Capture Tool, IDCT [44]. A set of attributes were measured and linked to each image: built plan area and number of stories, obtained from the cadastral map; and the type of lateral load-resisting system and the material and ductility of the lateral load-resisting system, assigned by visual inspection of the image. This task required 342 working hours, in a four months period, from a master student and two final year students of civil engineering. To guarantee a good performance on the assignation of the lateral load-resisting type, the students were trained on the guide for rapid visual screening of buildings for potential seismic hazards, published by Federal Emergency Management Agency, FEMA, [45].

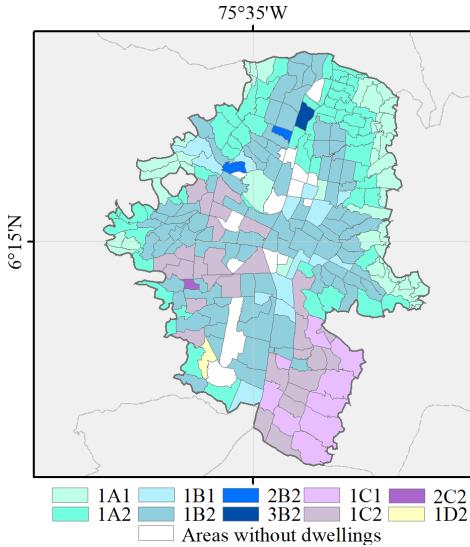


Fig. 3. Medellín division on homogeneous zones.



Fig. 4. Difficulties on visual building classification. (a) and (b): facade covering the structural system; (c) and (d): visual obstacles. Images source: photos taken by the authors in the city of Medellin. Jan., 2018

Two main difficulties aroused at the moment of building classification: structural system may be covered by the facade with different materials such as painting, masonry, plaster or stones (Fig. 4a,b); and visual obstacles such as vegetation and exterior structures (protection bars, walls, etc.) can be encountered (Fig. 4c,d). To overcome those situations the expert has two options: (1) visualize the building from different angles; and (2) assign a building classification based on the typology of the surrounding structures. Similar difficulties were encountered when collecting data for the development of exposure models of residential structures in Chile [9].

IV. EXPERIMENTAL SETUP

A total of 9,837 buildings (from the 11,381 surveyed buildings [16]) were considered for the Machine Learning process. Excluded buildings are those that have identical characteristics, as those buildings that belong to the same residential complex. Two macro classes were considered: masonry (M) buildings and reinforced concrete (RC) buildings. Fig. 5 presents the geographical distribution of the 9,837 buildings used for machine prediction. The spatial distribution of the macro classes shows three clear patterns: a high concentration of masonry in the north part of the city; a high concentration of reinforced concrete in the south-east part of the city; and a mixture of both macro classes in the rest of the city.

We used a finetuning strategy where we took publicly available pretrained CNNs, adapted the last fully connected layer

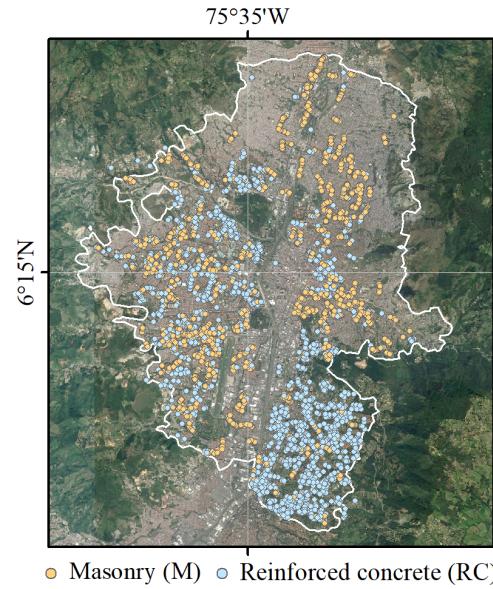


Fig. 5. Ground truth macro class distribution.

to our problem and ran a full training process with our data. We used AlexNet [19] a CNN winner of the 2012 ILSVRC [46] competition consisting on 5 convolutional layers followed by 3 fully connected, and VGG-16 [47], with 13 convolutional layers and 3 fully connected. Additionally, we established a baseline classification with random forests (RF) directly on the pixel images. We explored different RF parametrizations and established the baseline with the best one obtained. The experiments were run on the EAFIT Apolo cluster with a single process running on an NVIDIA K80 GPU with 64 GB RAM available (and 12GB on the GPU card).²

With this, our aims were: (1) to measure the actual benefit of using CNNs over classical machine learning techniques; and (2) understand the behavior of different levels of CNN complexity and assess strategies to improve the results.

With the annotated Google Street View images, we devised two experimental setups. In the first one, data was randomly split in 70% for training the model and 30% for measuring its classification accuracy. Then, we ran a 50/50 split twice, swapping train and test data so that we could obtain a classification assessment of the full dataset. This was done so that experts could manually assess CNN accuracy with the full dataset. Finally, the classification performance was measured in terms of accuracy (percentage of building images whose macro-class was correctly predicted by the system) and results are then discriminated by class.

V. RESULTS AND DISCUSSION

Table II shows the overall accuracy obtained in macro class prediction by the different experimental setups. We show both accuracies in the train and test data splits. The baseline was obtained by exploring some 50 RF configurations and the best baseline was selected which is what is shown. Observe that (1) CNNs do provide increased accuracy over this baseline; (2) the

²www.eafit.edu.co/apolo

increased complexity of VGG16 does not result in increased test accuracy; and (3) VGG16 does show overfitting (better train accuracy) which suggests possibly increased accuracy in test data with larger or cleaner datasets.

TABLE II
SUMMARY OF CNN MACRO CLASS PREDICTIONS

train/test split method	70/30		50/50	
	train	test	train	test
baseline (RF)	79%	80%	81%	78%
Alexnet	85%	85%	84%	85%
VGG16	95%	84%	91%	85%

Detailed CNN prediction performance is shown in Table III in terms of classification into the correct macro class. On the one hand, an excellent agreement was obtained in the masonry (M) macro class: 94% of the masonry buildings had a good prediction and only 6% were mispredicted. On the other hand, the percentage of successful predictions drops to 67% when the reinforced concrete (RC) macro class is considered.

TABLE III
DETAILED CNN PREDICTIONS

Macro class	Micro class	Prediction Correct	Incorrect	Total
Masonry (M)	MUR/LWAL/DNO	5,717 (95%) ^a	329 (5%) ^a	6,086 (92%) ^b
	MR/LWAL/DUC	149 (77%) ^a	45 (23%) ^a	194 (3%) ^b
	MFC/LWAL/DUC	135 (81%) ^a	32 (19%) ^a	167 (3%) ^b
	MFC/LWAL/DNO	142 (88%) ^a	20 (12%) ^a	162 (2%) ^b
	Total	6,183 (94%)^b	426 (6%)^b	6,609 (67%)^c
Reinforced concrete (RC)	CR/LINF/DNO	941 (49%) ^a	976 (51%) ^a	1,917 (60%) ^b
	CR/LINF/DUC	1,000 (93%) ^a	79 (7%) ^a	1,079 (33%) ^b
	CR/LWAL/DUC	121 (95%) ^a	7 (5%) ^a	128 (4%) ^b
	CR/LDUAL/DUC	104 (100%) ^a	0 (0%) ^a	104 (3%) ^b
	Total	2,166 (67%)^b	1,062 (33%)^b	3,288 (33%)^c
Total		8,369 (85%)^c	1,488 (15%)^c	9,837 (100%)^c

^a Percentage related to micro class

^b Percentage related to macro class

^c Percentage related to the total number of surveyed buildings

It is important to consider which errors under/over-estimate the exposure model. It can be said that an exposure model for seismic risk assessment has effect in two main components: structural behavior and value of the exposed assets. The main error that overestimates the structural performance in terms of structural behavior corresponds to the masonry buildings (M) classified as reinforced concrete (RC) —RC buildings are expected to have an overall better response to seismic load than M buildings. This error has a very small percentage of 6% of the masonry buildings, as shown in Table III (4% of the total number of buildings), which is highly acceptable. Underestimation of structural behavior takes place when RC buildings are classified as M buildings, which corresponds to 33% of the RC buildings (11% of the total number of buildings); the underestimation of structural behavior leads to conservative results in terms of number of damaged buildings and potential human losses.

In an exposure model development structural cost can be associated only to built-up area and socio-economic strata [2], [11], [12], [16]; in this situation the misclassification of buildings does not affect the structural cost of the exposed

assets. Nonetheless, if structural cost is function of building typology, overestimation of building cost takes place when M buildings are classified as CR (6% of M buildings; 4% of total number of buildings); whereas underestimation of building cost takes place when CR buildings as considered as M buildings (33% of CR buildings; 11% of total number of buildings). This error represents economical underestimation, but it is conservative in terms of structural performance.

TABLE IV
SUMMARY OF CNN MACRO CLASS PREDICTIONS PER HOMOGENEOUS ZONE

Macro class	Homogeneous zone ^a	Prediction Correct	Incorrect	Total ^b
Masonry (M)	1A1	242 (97%)	8 (3%)	250 (3.8%)
	1A2	827 (95%)	39 (5%)	866 (13.1%)
	1B1	535 (94%)	35 (6%)	570 (8.6%)
	1B2	3,622 (94%)	231 (6%)	3,853 (58.3%)
	2B2	45 (100%)	0 (0%)	45 (0.7%)
	3B2	15 (79%)	4 (21%)	19 (0.3%)
	1C1	104 (86%)	17 (14%)	121 (1.8%)
	1C2	733 (89%)	89 (11%)	822 (12.4%)
	2C2	10 (100%)	0 (0%)	10 (0.2%)
	1D2	50 (94%)	3 (6%)	53 (0.8%)
Reinforced concrete (RC)	1A1	12 (44%)	15 (56%)	27 (0.8%)
	1A2	70 (59%)	49 (41%)	119 (3.7%)
	1B1	119 (64%)	66 (36%)	185 (5.7%)
	1B2	807 (56%)	643 (44%)	1,450 (%)
	2B2	2 (67%)	1 (33%)	3 (0.1%)
	3B2	9 (82%)	2 (18%)	11 (0.3%)
	1C1	333 (81%)	77 (19%)	410 (12.7%)
	1C2	779 (79%)	206 (21%)	985 (30.5%)
	2C2	18 (95%)	1 (5%)	19 (0.6%)
	1D2	17 (89%)	2 (11%)	19 (0.6%)

^a First digit: predominant building height; middle digit: predominant socio-economic strata; last digit: remaining building height

^b Percentage related to macro class

Table IV presents the percentage of accurate predicted data for both macro classes, according to homogeneous zones (see code definition in Section III). Fig. 6 presents the percentage of buildings (bars) and number of buildings (numbers on top of bars) within each homogeneous zone that were classified as M or RC macro class, as well as building classification as function of number of stories. Bar width is associated with the number of surveyed buildings. It can be observed that, with the exception of two homogeneous zones, machine classification overestimates the number of M buildings, as the total number of buildings classified as M overpass the real number of M buildings (percentage of M buildings above 100%) (Fig. 6a). Overestimation is notorious in homogeneous zones 1C1 and 1C2, which are high income zones (in which it is expected to find better quality buildings). Underestimation of M buildings takes place at homogeneous zones 1D2 and 3B2, which are the more heterogeneous zones in terms of socio-economic strata and building number of stories. Underestimation of the number of RC buildings (percentage of RC buildings less than 100% on Fig. 6b) takes place in all of the zones with the exception of zone 3B2. Underestimation of RC buildings is more notorious in homogeneous zones with building predominant height of 1 and 2 stories and low to medium socio-economic strata (1A1, 1A2, 1B1 and 1B2). With the exception of 20 buildings, all of the mispredicted buildings of the dataset are low-rise

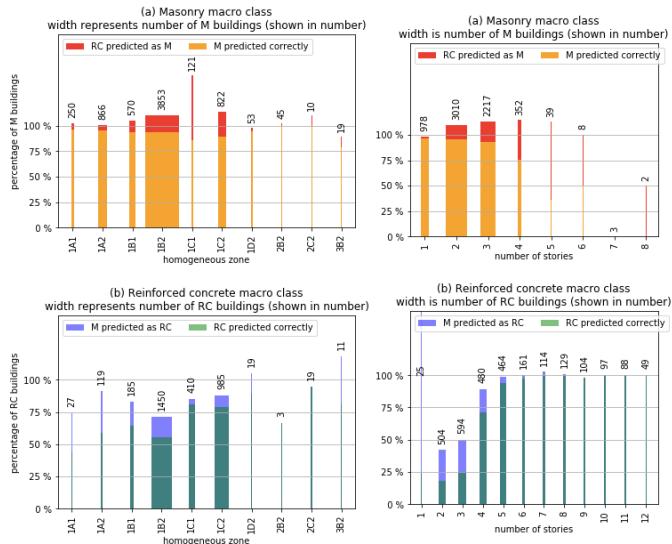


Fig. 6. Distribution of predicted buildings according homogeneous zones (left) and number of stories (right).

buildings with 1 to 5 stories. Analysis of results indicates that underestimation of RC buildings takes place at low socio-economic strata comunas, whilst overestimation of M buildings takes place at high socio-economic strata comunas.

Useful information is obtained when exploring the building classification of the macro classes' subdivision, i.e., the micro classes, as shown in Table III. If a threshold of 80% is considered for good/bad classification, it can be observed that for M macro class only the MR/LWAL/DUC (ductile reinforced masonry) micro class has a number of agreements smaller —although really close— to the given threshold. For the RC macro class, the CR/LINF/DNO (non-ductile RC infilled-frame) has an agreement of only 49%, whereas the remaining micro classes have an agreement above 90%. It can be concluded that the Machine Learning process have difficulties in the characterization of the MR/LWAL/DUC and CR/LINF/DNO micro classes. This misclassification can be considered as an acceptable mistake as (1) the MR/LWAL/DUC micro class corresponds to the masonry typology that has the best behavior under earthquake loading; and (2) CRLINF/DNO buildings are non-ductile, which implies the poorest performance under earthquake loading of the RC micro classes.

Fig. 7 shows the distribution of correct/incorrect predictions for the two aforementioned micro classes as function of the homogeneous zones and the number of stories. Percentage values (bars) must be interpreted carefully as number of sampled buildings (value in top of each bar) is reduced in some conditions. For MR/LWAL/DUC micro class it can be observed that the homogeneous zones with percentage of correct classification smaller than 75% and a relevant number of surveyed buildings are zones 1A1, 1A2, and 1B1. On the other hand, the number of stories with percentage of correct classification smaller than 75% and a relevant number of surveyed buildings is 4 and 5 stories. For CR/LINF/DNO micro class errors concentrate in a wider number of homogeneous zones: 1A1, 1A2, 1B1, 1B2, 1C1, and 1C2, and buildings with

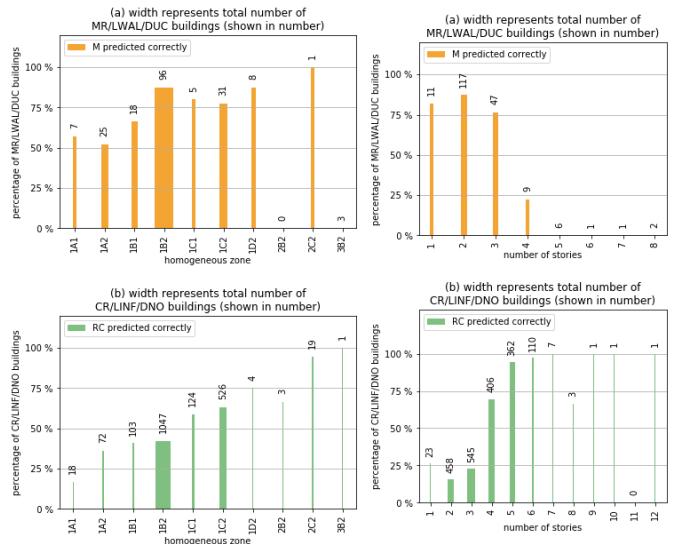


Fig. 7. Prediction distribution for MR/LWAL/DUC and CR/LINF/DNO according to homogeneous zones (left) and number of stories (right).

one to 4 stories.

Machine classifiers tend to favor majority classes [48]. Table V shows number of stories distribution for small rise buildings (1 to 4 stories) according to the considered macro classes. It can be observed that for 1 to 3 stories buildings the database has a small percentage of RC buildings, which may introduce a bias into the classification. Results from Machine Learning classifications could benefit if the number of MR/LWAL/DUC and CR/LINF/DNO surveyed buildings is increased, mainly on the homogeneous zones and number of stories previously described. Nonetheless, it must be kept in mind that visual differentiation between MR/LWAL/DUC and CR/LINF/DNO can be a challenge as the masonry reinforcement in MR/LWAL/DUC is not exposed and, in addition, it is common to cover the RC concrete structure of CR/LINF/DNO with materials such as masonry. Fig. 8 shows images of both MR/LWAL/DUC and CR/LINF/DNO in which even a structural expert can get confused. This situation could be partially solved if the geographical location is considered and the Machine Learning algorithm includes a prioritization scheme based on the characteristics of the zone (such as socio-economic strata).

TABLE V
DISTRIBUTION OF BUILDING MACRO CLASSES FOR LOW-RISE BUILDINGS

Macro class	Number of stories			
	1	2	3	4
Masonry (M)	978 (89%)	3,010 (86%)	2,217 (79%)	352 (42%)
Reinforced concrete (RC)	25 (2%)	504 (14%)	594 (21%)	480 (58%)
Total	1,003	3,514	2,811	832

VI. CONCLUSIONS AND FUTURE WORK

As a final remark, it must be kept in mind that building typology distribution of the presented database does not represent building distribution in the city of Medellín; it represents

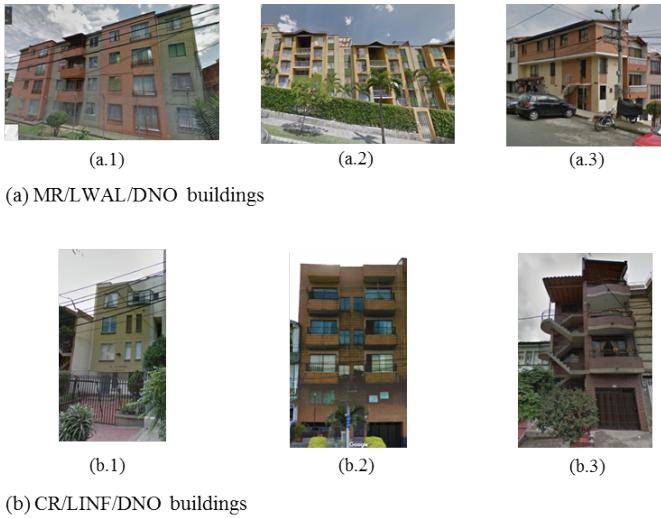


Fig. 8. Examples of buildings in which differentiation between MR/LWAL/DUC and CR/LINF/DNO micro classes can be difficult. Images source: c 2018 Google "Street View" digital images, Google Maps (<http://maps.google.com>), image capture: (a.1) Jul. 2017; (a.2) Sep. 2015; (a.3) May 2013; (b.1 and b.2) Aug. 2017; and (b.3) Dec. 2013.

the building typology distribution of the surveyed buildings. Results from the Machine Learning process are intended to be used by the exposure model developer in the task of gathering statistical information that must be processed in order to assign building typology distribution to non-surveyed buildings [12], [16].

The methodology described in this paper can be applied to other regions by considering that building typology varies from place to place. Different building typologies must be known a priori in order to define macro/micro classes. In those areas in which imagery dataset is reduced, local expert judgment as well as Bayesian networks strategies can increase the machine performance.

Micro class classification is highly valuable information for exposure model developers. Future improvements should focus on micro class classification. An additional desirable output would be the identification of building number of stories from the Google Street View images.

Better machine performance is expected if information from facade images is complemented with data from aerial photography and parameters such a building number of stories are included in the machine learning process. In addition, future work will include the influence of: (1) under/over-estimation of building classification in a seismic risk assessment, and (2) human level of accuracy on building typology classification.

REFERENCES

- [1] C. Geiß and H. Taubenböck, "Remote sensing contributing to assess earthquake risk: from a literature review towards a roadmap," *Natural Hazards*, vol. 68, no. 1, pp. 7–48, 2013.
- [2] C. Yépez-Estrada, V. Silva, J. Valcárcel, A. B. Acevedo, N. Tarque, M. A. Hube, G. Coronel, and H. Santa María, "Modeling the residential building inventory in south america for seismic risk assessment," *Earthquake Spectra*, vol. 33, no. 1, pp. 299–322, 2017.
- [3] "Hazus-mh mr5, multi-hazard loss estimation methodology–earthquake model," Department of Homeland Security, Federal Emergency Management Agency, Mitigation Division, Washington, DC, USA, Tech. Rep., 2010.
- [4] V. Silva, H. Crowley, M. Pagani, D. Monelli, and R. Pinho, "Development of the openquake engine, the global earthquake model's open-source software for seismic risk assessment," vol. 72, 07 2014.
- [5] F. Yamazaki, H. Mitomi, M. Matsuoka, and K. Honda, "Inventory development for natural and built environments-remote sensing technologies for inventory development and risk assessment-characteristics of satellite images in bangkok, thailand," NIED, Miki, Hyogo, Japan, Tech. Rep. on The Development of Earthquake and Tsunami Mitigation Technologies and their Integration for the Asia-Pacific Region, Tech. Rep., 2000.
- [6] A. C. Caputo and A. Vigna, "Numerical simulation of seismic risk and loss propagation effects in process plants: An oil refinery case study," in *ASME 2017 Pressure Vessels and Piping Conference*. American Society of Mechanical Engineers, 2017, pp. V008T08A024–V008T08A024.
- [7] M. Pittore and M. Wieland, "Toward a rapid probabilistic seismic vulnerability assessment using satellite and ground-based remote sensing," vol. 68, pp. 115–145, 2013.
- [8] G. C. Iannelli and F. Dell'Acqua, "Extensive exposure mapping in urban areas through deep analysis of street-level pictures for floor count determination," vol. 1, no. 16, 2017.
- [9] H. Santa María, M. A. Hube, F. Rivera, C. Yépes-Estrada, and J. A. Valcárcel, "Development of national and local exposure models of residential structures in chile," *Natural Hazards*, vol. 86, no. 1, pp. 55–79, 2017.
- [10] M. Pittore, M. Wieland, and K. Fleming, "Perspectives on global dynamic exposure modelling for geo-risk assessment," *Natural Hazards*, vol. 86, no. 1, pp. 7–30, 2017.
- [11] M. A. Salgado-Gálvez, D. Zuloaga-Romero, G. A. Bernal, M. G. Mora, and O.-D. Cardona, "Fully probabilistic seismic risk assessment considering local site effects for the portfolio of buildings in medellín, colombia," *Bulletin of Earthquake Engineering*, vol. 12, no. 2, pp. 671–695, 2014.
- [12] A. B. Acevedo, J. D. Jaramillo, C. Yépes-Estrada, V. Silva, F. A. Osorio, and M. Villar-Vega, "Evaluation of the seismic risk of the unreinforced masonry building stock in antioquia, colombia," *Natural Hazards*, vol. 86, no. 1, pp. 31–54, 2017.
- [13] D. Dutta and K. Serker, "Urban building inventory development using very high-resolution remote sensing data for urban risk analysis," *International Journal of Geoinformatics*, vol. 1, no. 1, pp. 190–116, 2005.
- [14] H. Taubenböck, A. Roth, and S. Dech, "Vulnerability assessment using remote sensing: The earthquake prone megacity istanbul, turkey," *Proceedings of ISRSE 2007*, pp. 1–5, 2007.
- [15] S. Valero, J. Chanussot, and P. Gueguen, "Classification of basic roof types based on vhr optical data and digital elevation model," in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 4, 2008, pp. IV – 149–IV – 152.
- [16] D. Gonzalez and A. B. Acevedo, "Actualización del modelo de exposición sísmica para viviendas en medellín (colombia) y su aplicación en la evaluación del riesgo sísmico para viviendas de mampostería no reforzada," in *VIII Congreso Nacional de Ingeniería Sísmica*, Barranquilla, Colombia, 2017.
- [17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [18] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [20] S. Esposito, I. Iervolino, A. d'Onofrio, A. Santo, F. Cavalieri, and P. Franchin, "Simulation-based seismic risk assessment of gas distribution networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 30, no. 7, pp. 508–523, 2015.
- [21] D. Cook, K. Fitzgerald, T. Chrupalo, and C. Haselton, "Comparison of fema p-58 with other building seismic risk assessment methods," 2017.
- [22] N. Frolova, V. Larionov, J. Bonnin, S. Sushchev, A. Ugarov, and M. Kozlov, "Seismic risk assessment and mapping at different levels," *Natural Hazards*, vol. 88, no. 1, pp. 43–62, 2017.
- [23] S. Mangalathu, J.-S. Jeon, J. E. Padgett, and R. DesRoches, "Performance-based grouping methods of bridge classes for regional seismic risk assessment: Application of anova, ancova, and non-parametric approaches," *Earthquake Engineering & Structural Dynamics*, 2017.

- [24] I. Riedel, P. Guéguen, M. Dalla Mura, E. Pathier, T. Leduc, and J. Chanussot, "Seismic vulnerability assessment of urban environments in moderate-to-low seismic hazard regions using association rule learning and support vector machine methods," *Natural Hazards*, vol. 76, no. 2, pp. 1111–1141, 2015.
- [25] Y. Zhang, H. V. Burton, H. Sun, and M. Shokrabadi, "A machine learning framework for assessing post-earthquake structural safety," *Structural Safety*, vol. 72, pp. 1–16, 2018.
- [26] W. D. Fisher, T. K. Camp, and V. V. Krzhizhanovskaya, "Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection," *Journal of Computational Science*, vol. 20, pp. 143–153, 2017.
- [27] Y. Bai, B. Adriano, E. Mas, H. Gokon, and S. Koshimura, "Object-based building damage assessment methodology using only post event alos-2/palsar-2 dual polarimetric sar intensity images," *Journal of Disaster Research Vol.*, vol. 12, no. 2, p. 259, 2017.
- [28] M. Wieland, M. Pittore, S. Parolai, and J. Zschau, "Exposure estimation from multi-resolution optical satellite imagery for seismic risk assessment," *ISPRS International Journal of Geo-Information*, vol. 1, no. 1, pp. 69–88, 2012.
- [29] C. Geiß, A. Schauß, T. Riedlinger, S. Dech, C. Zelaya, N. Guzmán, M. A. Huber, J. J. Arsanjani, and H. Taubenböck, "Joint use of remote sensing data and volunteered geographic information for exposure estimation: evidence from valparaíso, chile," *Natural Hazards*, vol. 86, no. 1, pp. 81–105, 2017.
- [30] A. Raouf, Y. Peng, and T. I. Shah, "Integrated use of aerial photographs and lidar images for landslide and soil erosion analysis: a case study of wakamow valley, moose jaw, canada," *Urban Science*, vol. 1, no. 2, p. 20, 2017.
- [31] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," *arXiv preprint arXiv:1702.02073*, 2017.
- [32] I. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks, 2014," *arXiv preprint Arxiv:1312.6082*, 2015.
- [33] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991–2000, 2014.
- [34] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei, "Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states," *Proceedings of the National Academy of Sciences*, p. 201700035, 2017.
- [35] "Estudio general de amenaza sísmica de colombia," Asociación Colombiana de Ingeniería Sísmica, Comité AIS-300, Bogotá, Colombia, Tech. Rep., 2009.
- [36] L. F. Restrepo, M. R. Villarraga, J. D. Jaramillo, Y. Farbizar, A. F. Vélez, D. A. Rendón, and F. P. Ángel, *Microzonificación y evaluación del riesgo sísmico del Valle de Aburrá*. Área Metropolitana del Valle de Aburrá, 2007.
- [37] "Instrumentación y microzonificación sísmica del área urbana de medellín," Sistema Municipal para la Prevención y Atención de Desastres, Medellín, Colombia, Tech. Rep., 1999.
- [38] K. Jaiswal and D. J. Wald, "Creating a global building inventory for earthquake loss assessment and risk management," U.S. Geological Survey Open-File Report 2008-1160, Tech. Rep., 2008.
- [39] C. Rojahn and R. L. Sharpe, "Earthquake damage evaluation data for california, atc-13," Applied Technology Council, Tech. Rep.
- [40] "Hazus-mh mr4 technical manual," Federal Emergency Management Agency, Mitigation Division, Washington, DC, USA, Tech. Rep., 2003.
- [41] S. Brzev, C. Scawthorn, A. W. Charleson, L. Alle, M. Greene, K. Jaiswal, and V. Silva, "Gem building taxonomy version 2.0," GEM Foundation, Pavia, Italy, Tech. Rep. 2013-02 V1.0.0, Tech. Rep., 2013.
- [42] "Reglamento colombiano de construcción sismo resistente," Asociación Colombiana de Ingeniería Sísmica, Bogotá, Colombia, Tech. Rep., 2010.
- [43] "Código colombiano de construcción sismo resistente," Asociación Colombiana de Ingeniería Sísmica, Bogotá, Colombia, Tech. Rep., 1984.
- [44] C. J. Jordan, K. Adlam, K. Laurie, W. Shelley, and J. Bevington, "User guide: windows tool for field data collection and managment," GEM Foundation, Pavia, Italy, Tech. Rep. 2014-04 V1.0.0, Tech. Rep., 2014.
- [45] "Rapid visual screening of buildings for potential seismic hazards: supporting documentation," FEMA 154, Washington, DC, USA, Standard, 2002.
- [46] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [48] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, "A supervised learning approach for imbalanced data sets," in *19th ICPR, Tampa, FL, USA*, 2008.



Daniela González received the degrees of Civil Engineer and MSc in Engineering from Universidad EAFIT, where she is currently completing a PhD in Engineering. She has worked as a consultant on structural design and analysis, and has been teaching structural analysis to undergraduate students since 2017. Her research interests focus on disaster mitigation and seismic risk assessment; she developed a seismic exposure model for the city of Medellín and co-authored exposure models for Bogotá and Cali, two of the main cities of Colombia.



Raúl Ramos-Pollán holds a professor position at the School of Computer Science at Universidad de Antioquia (Colombia). PhD in Informatics Engineering at University of Porto, Portugal, his career developed between industry and academy and includes the functions of Director of the CETA-CIEMAT Supercomputing Center (Trujillo, Spain), Director of the Software Department at Pildo Labs (aeronautics SME in Barcelona, Spain), Java Architect for Sun Microsystems (Geneve, CH), Software Engineer at CERN (Geneve, CH), Senior Researcher at Universidad Nacional de Colombia and professor at Universidad Industrial de Santander. His research interests are currently focused on image analytics and deep learning.



Ana B. Acevedo is a Full Professor in the Civil Engineering Department at Universidad EAFIT. She received the B.S. degree in Civil Engineering from the National University of Medellín, Colombia; and the M.S. and Ph.D degrees on Reduction of Seismic Risk from the ROSE School–University of Pavia (Italy). Her current research interests include seismic risk assessments, seismic structural performance and disaster mitigation.



Juan C. Duque is a Full Professor in the Department of Mathematical Sciences at Universidad EAFIT. Duque is PhD in Business Studies from University of Barcelona, and MSc in Economics and Business from Pompeu Fabra University. He is the founder/director of RISE group (Research in Spatial Economics) and member of the Geographic Systems Analysis Lab (GSAL) group of the University of California (Santa Bárbara). His research topics are Regional Science, Operation Research and Spatial Analysis. Duque is member of the editorial board of International Regional Science Review and Computers Environment and Urban Systems.



Alejandro Betancourt Ph.D. with Cum Laude in Cognitive Environments (2017) of the Technical University of Eindhoven (The Netherlands) and the University of Genova (Italy). Master in Applied Mathematics (2013) and Bachelor in Mathematical Engineering (2011) from Universidad EAFIT. Alejandro is currently working as Digital Transformation Leader in Ecopetrol and his research interests include Artificial Intelligence, Distributed Computing, and Corporative Digital Strategy.



Sebastian Garcia Valencia is MSc in Engineering from Universidad EAFIT University, with studies in the MSc in Computer Science of the TU Ilmenau and Bachelor in Computer Science from Universidad EAFIT. Sebastian has worked in the laboratories of IBM R&D (Germany), the HPC center Apolo, the RiSE Group and the Center of Excellence and Appropriation on Big Data and Data Analytics (CAOBA). His research interests include Machine Learning, Music DPS and Distributed Computing.



Diego Rueda-Plata received a bachelor's degree in Computer Systems Engineering and currently enrolled in a Master of Applied Mathematics at the School of Physics at Universidad Industrial de Santander (Colombia). Currently, he is working with deep learning using unmanned aerial vehicles and research on image analytics and deep learning.