Use of Sentinel 2 Images and Cadastral Data to Analyze Urban Growth in Colombia

Rodolfo Yesid MEZA PATACÓN and Sonia Constanza GARZÓN MARTÍNEZ, Colombia

Key words: urban growth, Sentinel 2, Artificial Intelligence, Data Analysis, Cloud Computing

SUMMARY:

The Institute of Geography Agustín Codazzi (IGAC) of Colombia stands at the forefront of cadastral innovation, leveraging geospatial technologies to bolster national development. The institute has embarked on a pivotal research initiative, the outcomes of which are encapsulated in the Physical Change Monitoring System. This system is bifurcated into two primary components: a) the forecasting of urban expansion, and b) the multi-temporal analysis of construction alterations. The project's objectives were twofold: a) to harness satellite imagery for the prognostication of urban sprawl patterns, and b) to scrutinize the precision and efficacy of Google's construction projections by juxtaposing them with historical cadastral data. The project's methodology unfolded over three distinct phases: 1) Data acquisition: This phase involved the collection of Sentinel 2 satellite imagery, the compilation of IGAC's cadastral database, and the assimilation of the Google Open Buildings database. 2) Big Data Processing: This phase saw the deployment of cloud computing resources to efficiently manage extensive datasets with geospatial attributes. 3) Machine Learning: This phase entailed the application of artificial intelligence algorithms with a dual aim—predicting urban expansion trajectories and evaluating the veracity of Google's structural models. Preliminary findings reveal a tangible correlation between Sentinel 2's imagery and concrete patterns of urban growth. Moreover, the disparities discerned between Google's forecasts and actual cadastral records furnish valuable insights, paving the way for the refinement of cadastral administration methodologies. However, the endeavor introduced significant challenges, most notably the need to scale the models to accommodate Colombia's diverse regions and the handling of copious amounts of geospatial data. Future directions involve the fine-tuning of models to adeptly reflect the unique geographical nuances inherent to the Colombian landscape. This meticulous approach ensures that the project's insights and methodologies are conveyed with the requisite clarity and precision befitting a technical paper, facilitating a comprehensive understanding of the initiative's scope, methodologies, and implications.

RESUMEN:

El Instituto Geográfico Agustín Codazzi (IGAC) de Colombia es líder en temas catastrales que aprovecha las tecnologías geoespaciales para promover el desarrollo del país, por lo que adelantó un proyecto de investigación con resultados en el Sistema de Monitoreo de Cambios Físicos desarrollado en dos componentes: a) La predicción del crecimiento urbano y b) Los cambios multitemporales de las construcciones. Entre los objetivos del proyecto se tuvieron: a) Utilizar imágenes para predecir patrones de crecimiento urbano. b) Evaluar la eficacia y precisión de las predicciones de construcción de Google comparándolas con los registros catastrales de épocas anteriores. El proyecto contempló tres etapas metodológicas que fueron: 1) Colecta de datos: Utilización de imágenes de satélite de Sentinel 2, Censo poblacional DANE 2) Procesamiento Big Data: Empleando servicios de computación en la nube para manejar conjuntos de datos de grandes dimensiones con componente geográfico, 3) Aprendizaje automático: Aplicación de algoritmos de inteligencia artificial para predecir el crecimiento urbano y evaluar la precisión de los modelos de construcción de Google. Los resultados iniciales indican una correlación entre las imágenes de Sentinel 2 y los patrones reales de crecimiento urbano. Además, las diferencias entre las predicciones de Google y los registros catastrales proporcionan información para mejorar las estrategias de gestión catastral. Con este proyecto se planteó el reto de escalamiento de los modelos a distintas regiones del país, así como el procesamiento de grandes volúmenes de información geográfica, siendo necesario a futuro la configuración de modelos que capten las diferencias regionales del territorio colombiano.

Use of Sentinel 2 Images and Cadastral Data to Analyze Urban Growth in Colombia

Rodolfo Yesid MEZA PATACÓN and Sonia Constanza GARZÓN MARTÍNEZ, Colombia

1. INTRODUCTION:

The Institute of Geography Agustín Codazzi (IGAC) of Colombia is the national cadaster authority in charge of the processes of formation, updating, conservation and dissemination of the multipurpose cadaster for the country. (Congress of Colombia, 2019). The cadaster is defined as the inventory or census of real estate located in the national territory, of public or private domain, which must be updated and classified to achieve its physical, legal, and economic identification. On the other hand, the cadaster with a multipurpose approach is one in which the information produced from its operation serves as input for the formulation and implementation of public policies, providing greater legal certainty, efficiency of the real estate market, development, and land use planning, among other benefits (DANE, 2020).

Since 2016, the elevated level of outdated cadaster in Colombia was diagnosed, finding by that time that only 320 municipalities out of 1101 of the national totals had a status of the cadastral update (CONPES, 2016). As a result of pilot exercises in 11 municipalities in the country, regarding the implementation of multipurpose cadaster, it was identified that the inputs for field work are limited or deficient, which hinders and delays the cadastral survey in a territory; therefore, by means of the (CONPES, 2019), line of action 12 was defined regarding the Identification, development, and adoption of updating schemes and more efficient technologies to keep the cadastral information updated, where the IGAC is assigned the task of reviewing the technologies used in the country or outside it, for the identification of the expansion of urban boundaries or changes in properties or constructions, these as indicators of the dynamics of the territory in cadastral matters.

As a result of the above need, two projects were advanced in 2020 and 2022 by the IGAC Research Center, which advanced in results in relation to the exploration of artificial intelligence models for the classification of high-resolution images regarding the level of constructions and the comparison of constructions obtained in two different periods, to identify possible changes; on the other hand, the review of a convolutional neural network model and the corresponding predictor variables of the urban expansion phenomenon. The above advances were made in study areas of the country, where the availability of information inputs was favorable.

By the year 2022, and with the implementation of the new National Development Plan for Colombia (2022 - 2026), challenging goals associated with the Multipurpose Cadaster are imposed on the IGAC, where it is necessary to go from a 9.4% update rate in 2022 to 70% by 2026. (Congress of Colombia, 2023). Given the above situation, the IGAC resolves from the research results obtained previously, to think of a technological solution that will integrate the components of change of constructions in two dimensions and estimates of urban expansion, this bearing in mind the following technical objectives (For the purposes of this document, progress in the development of the Physical Change Monitoring System, associated with the urban expansion component, is shown):

- Evaluate, adapt and scale the results of the models generated to different regions of the country, especially in those areas prioritized by the cadastral updating process.
- Take advantage of institutional and open-source data that allow implementing the new technological solution as part of the productive process of the cadastral management advanced by the IGAC and define a process of extraction, transformation and ETL loading linked to the different data sets.
- Define the architecture of a system for monitoring physical changes for cadastral purposes, capable of modularly coupling the different components that make up the indicators of change in the territory.

2. METHODOLOGY:

The study was developed considering key aspects of data science, appropriate data capture and transformation, model training, model validation and finally a deployment so that different users can access the model results. Some relevant aspects that give context to the distinct phases are presented below.

2.1. Area of study:

The initial study, focused on the municipality of Chía in the department of Cundinamarca (See Figure 1), was a crucial starting point in the application of advanced remote sensing techniques to understand urban sprawl in Colombia. With the support of the Institute of Geography Agustín Codazzi (IGAC), this project has prioritized the inclusion of specific municipalities, adapting its methodologies to the unique needs and characteristics of each area. In a later phase, the study was significantly extended to the department of Meta, initially incorporating six municipalities in the analysis (See Figure 1).

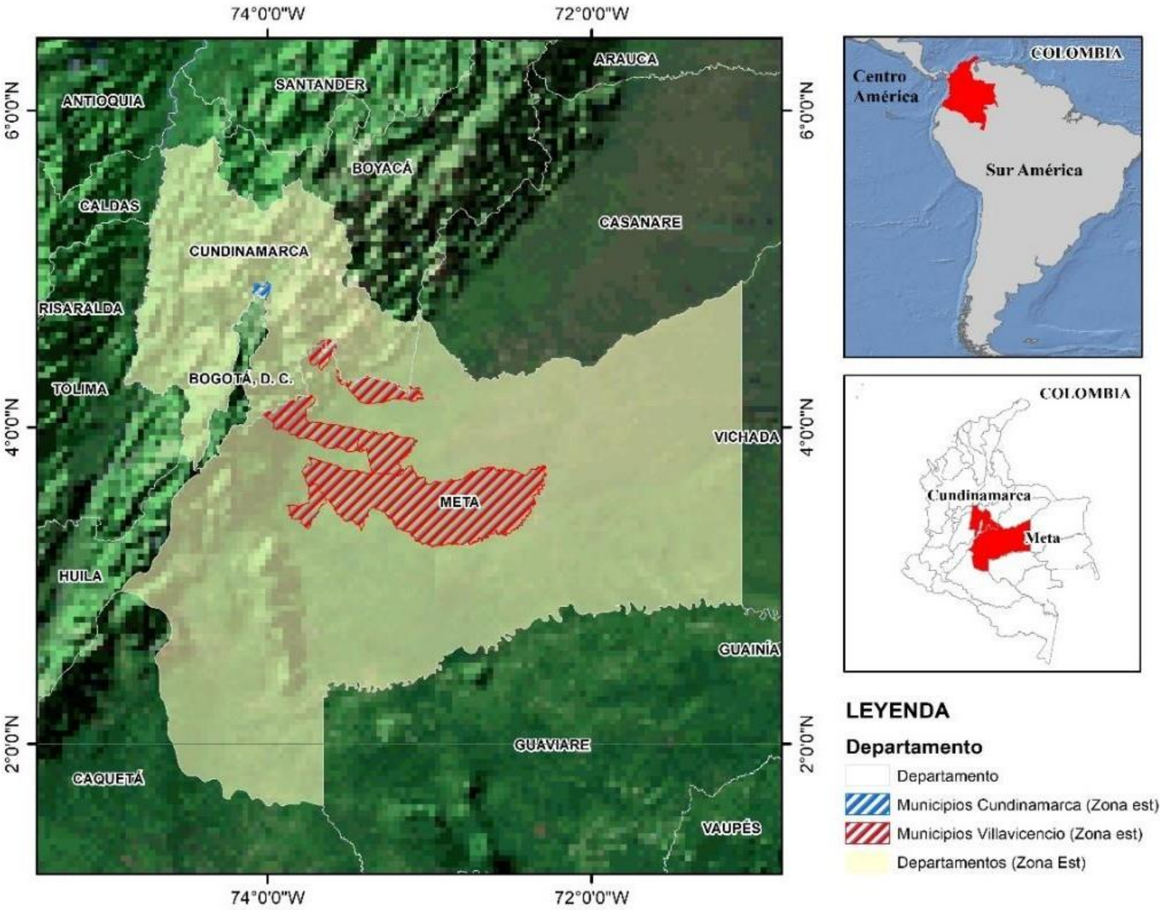

Figure 1. Location of the pilot study area for project 2022 Cundinamarca - Implementation 2023 Meta Municipalities.

This expansion represented a challenge in terms of data processing and analysis, providing a unique opportunity to apply and refine the methodologies in a broader and more diverse context. Some model trainings were carried out to scale up to 49 municipalities; however, due to the diversity of the municipalities included in this last phase, the need to optimize the datasets and segment the study area was identified, considering municipalities that have similarities in terms of reflectance levels measured from Sentinel 2 images. (See Figure 2).

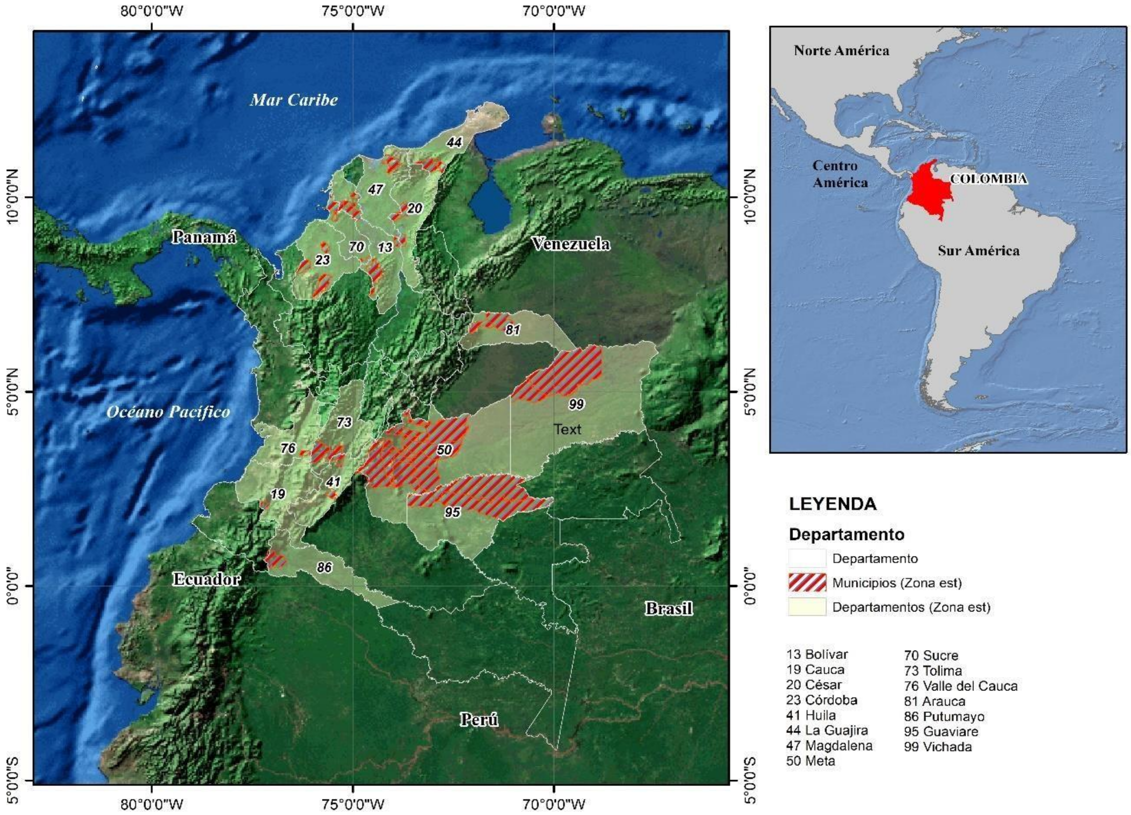

Figure 2. Location of the 49 municipalities considered during the training of the models
Source: Own elaboration.

2.2. Inputs:

An essential part of the study was the selection and processing of data for training the remote sensing model. This process was carried out in several stages, each focused on ensuring the accuracy and relevance of the data used.

- Sentinel-2 imagery: The basis of the analysis was the Sentinel-2 satellite imagery acquired through the Google Earth Engine API. These images are fundamental for the identification and monitoring of land cover changes and urban sprawl.

- Sentinel-2 Image Preprocessing: The areas corresponding to each block were extracted from Sentinel-2 satellite imagery, with special attention to the 2018 census data. This step was crucial to ensure that the model was trained with geographically accurate and up-to date information.
- DANE 2018 Census Data: Another key component in the model was the 2018 Census data provided by the National Administrative Department of Statistics (DANE). This data

included key information such as population in each municipality and building density at the block level, arranging a geo-referenced dataset. This data was processed to create raster images with a 10m resolution at the pixel level, which allowed for detailed and accurate integration with Sentinel II satellite imagery.

The use of these inputs, combined with advanced data processing techniques, allowed the creation of a robust and reliable model for the analysis of urban sprawl. This multidimensional approach not only improved the accuracy of the model, but also ensured that the results were relevant and applicable to the Colombian context.

2.3. Tools:

The urban expansion study in the department of Meta, Colombia, relied on a diverse set of technological tools for data processing and model training. These tools facilitated large-scale data analysis and allowed for greater accuracy and efficiency throughout the process.

- **Google Earth Engine (GEE):** The Google Earth Engine API was used to automate the download of Sentinel-2 satellite images. This process included the download of specific urban areas of each municipality, using centralized coordinates and a defined radius of action.
- **Python and Jupyter Notebooks:** The Python programming language was essential to the study, especially in the data processing phase. Numerous Python scripts and notebooks were developed and used to handle complex tasks such as image extraction, coordinate calculation, and dependent variable processing for the training model.
- **Integration and Scalability:** In addition to using these tools for data processing and model training, we also focused on the integration and scalability of the solution. This included optimizing the raster download of the urban sprawl estimation for each municipality and compatibility with different Python frameworks and Jupyter Notebooks.

The use of Jupyter Notebooks provided an interactive and user-friendly platform for experimentation and data visualization. The integration of ipywidgets into the notebooks further enhanced the interactivity and user experience, allowing for more dynamic and accessible data analysis.

2.4. Step by step:

The study on urban sprawl in the department of Meta, Colombia, followed a structured and detailed methodology, from the initial data processing to the final deployment of the visualizations (see Figure 3. Methodological flow of the project).

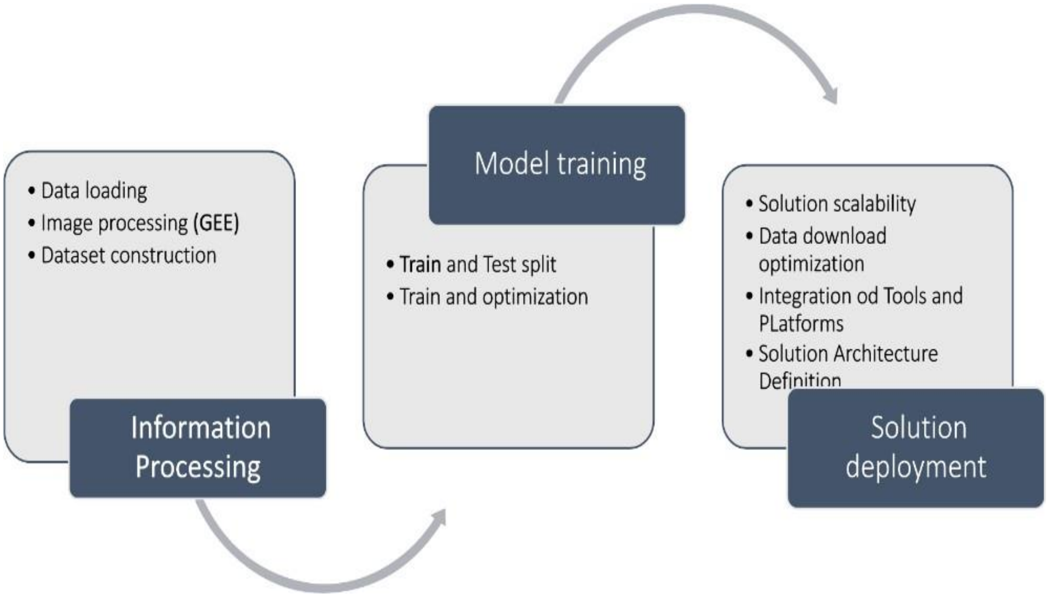

Figure 3. Methodological flow of the project. Source: Own elaboration.

Information processing:

- **Data Loading:** Initially, we loaded the data sets corresponding to the municipalities studied. This included data from six municipalities and was later expanded to 49 municipalities.
- **Image Processing in Google Earth Engine (GEE):** Google Earth Engine was used to download images from the Sentinel 2 project through its Python API. Zones of 2.5 km around the midway point of the municipality were defined to perform the download.
- **Tags:** The 2018 DANE Census was taken as a reference to extract the information associated with the characteristics of urban behavior in terms of number of dwellings, population, and density per urban block.
- **Dataset Construction:** Datasets for the model were constructed by reading image folders and labels, obtaining indexes, and creating data arrays in Numpy format.

Model Training:

- **Separation of Train and Test:** A careful separation of the data into training and test sets was performed. This allowed validating the accuracy and generalization of the model.
- **Training and Optimization:** The model was trained using the processed data, maintaining hyperparameters and the convolutional neural network structure. Different training strategies were considered, considering data size and computational capacity. The best

results were obtained using a RESNET 50 architecture implemented with Keras and Tensorflow.

Solution Deployment:

- **Scalability of the Solution:** A key aspect was scalability, considering the growing number of municipalities and the integration of additional data.
- **Data Download Optimization:** We optimized the download of urban growth estimation rasters for each municipality, facilitating access through web browsers.
- **Integration of Tools and Platforms:** The solution was integrated with a Python framework and made compatible with ipywidgets available in Jupyter Notebooks, allowing efficient control of permissions for users and resources.
- **Definition of the solution architecture**: Which follows the guidelines of the ISO 42010 standard, which develops the stages of analysis and design of the technological solution

2.1 RESULTS:

The results are presented in a manner consistent with the objectives of the project, as follows:

Definition of Extract, Transform and Load (ETL) process:

The quality and relevance of the data used in the training plays a key role in the results, in this case the labels are calculated from information available from the DANE 2018 Census selected for its inherent relationship to urban growth (Aspinall, R, 2004). Considering the presence of missing data in each of these characteristics, a composite indicator is constructed that involves proportionally each of the selected census characteristics and reduces the possibility of assigning labels with null values in the training phase (See equation in Figure 4).

The construction of this labeled dataset is done from Python scripts, including data balancing schemes, to ensure a balance between urban and rural samples prior to the training process. The final data sets were constructed with Numpy array format, ensuring the elimination of null data and the proper standardization of values between 0 and 1 to optimize the learning process and the convergence of the neural network. These arrays are constructed from tessellations associated with 13x13px regions, with center at the pixel labeled with the ICU, following Tobler's first law (Sui, D, 2004).

$$ICU = \sum_{J=1}^{pxmanzana} k1 * POBLACION + K2 * VIVIENDA + K3 * \ DENSIDAD$$

Figure 4. Composite urban growth index.

The construction of this labeled dataset is done from Python scripts, including data balancing schemes, to ensure a balance between urban and rural samples prior to the training process. The final data sets were constructed with Numpy array format, ensuring the elimination of null data and the proper standardization of values between 0 and 1 to optimize the learning process and the convergence of the neural network. These arrays are constructed from tessellations associated with 13x13px regions, with center at the pixel labeled with the ICU, following Tobler's first law (Sui, D, 2004).

Evaluation, adaptation, and scaling of the model:

In the preliminary research phase, model training was performed for the municipality of Chia in Colombia, this phase showed the feasibility of implementing convolutional neural networks for the detection of urban growth from satellite images with 10m resolution, however, the study area was limited, some steps of data set construction were performed manually with the help of GIS software and Python scripts with iterative functions, with a moderate computational cost. One of the objectives of this new phase was to achieve scaling, not only of the data processing, but also of the training on data sets corresponding to different municipalities to generalize the model.

In the processing stage, some programming routines were included to speed up the download of information with the Python API for Google Earth Engine. In training, the locally available computational capacity was increased, particularly RAM memory and GPU usage. Additionally, data augmentation processes were taken into account, which are necessary to achieve a better generalization of the model and what this entails in terms of the volume of information. (Zhang, K; Xu, G; Han, Z, 2020). The result with a data set of approximately 10GB corresponding to 6 municipalities of the department of Meta, is shown in Figure 5 - Response of the Convolutional Neural Network Model (CNN). An adequate learning behavior is evidenced with the training set and in turn a response with a similar trend with the test data, although of course with a much higher loss response.

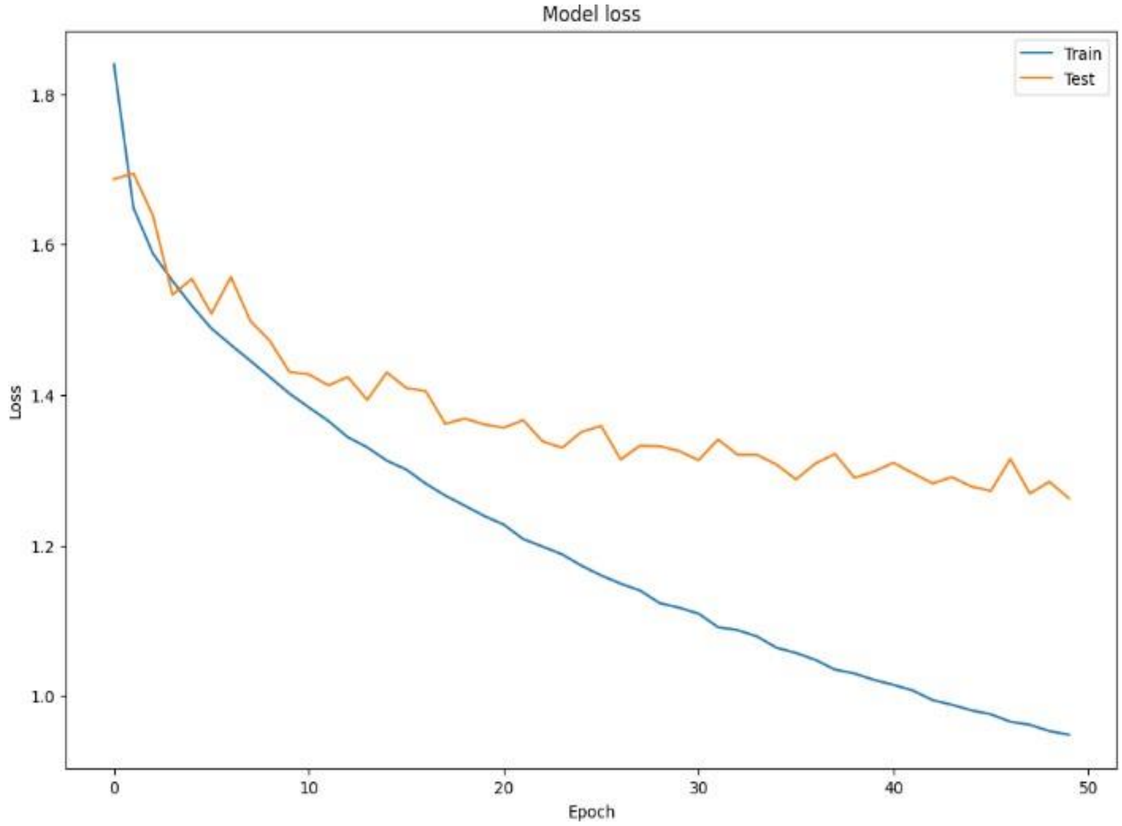

Figure 5. CNN model response.

In real terms, the prediction made by the model when given Sentinel 2 images from years after the 2018 benchmark, shows an adequate behavior in the recognition of the urban footprint of the municipality. In some areas that correspond to water bodies with sedimentation the model fails to adequately interpret the difference, eventually this may be due to the way the data balancing was generated and that this type of pixels was mostly eliminated during this process. Figure 6 - visual results of the urban area reconnaissance, shows not only the estimate made by the model for the urban area, in this case the image reflects the difference in UCI (Urban Growth Index) values between 2022 and 2019.

The areas in red indicate a greater change in the levels calculated from the reflectance levels present in the Sentinel 2 images. For some municipalities, the difference in UCI was contrasted with multitemporal images of higher resolution; it was possible to verify that in cases of greater intensity of red a key component of the variation was new constructions or changes in existing constructions. However, it is important to note that not all the change is due to the identification of variations in constructions, other factors may affect, for example, the average reflectance for image collection, permissible cloudiness level established at the time of discharge or minor changes in the roofs of the properties, so it is necessary that the cadaster analyst takes into account these aspects when making decisions on the level of change attributable exclusively to urban growth.

Physical change monitoring system architecture:

The monitoring system we have developed to study physical changes in urban areas in Colombia is structured in two main blocks: information processing and model deployment. This architecture is designed to efficiently capture, process, and visualize data related to urban sprawl.

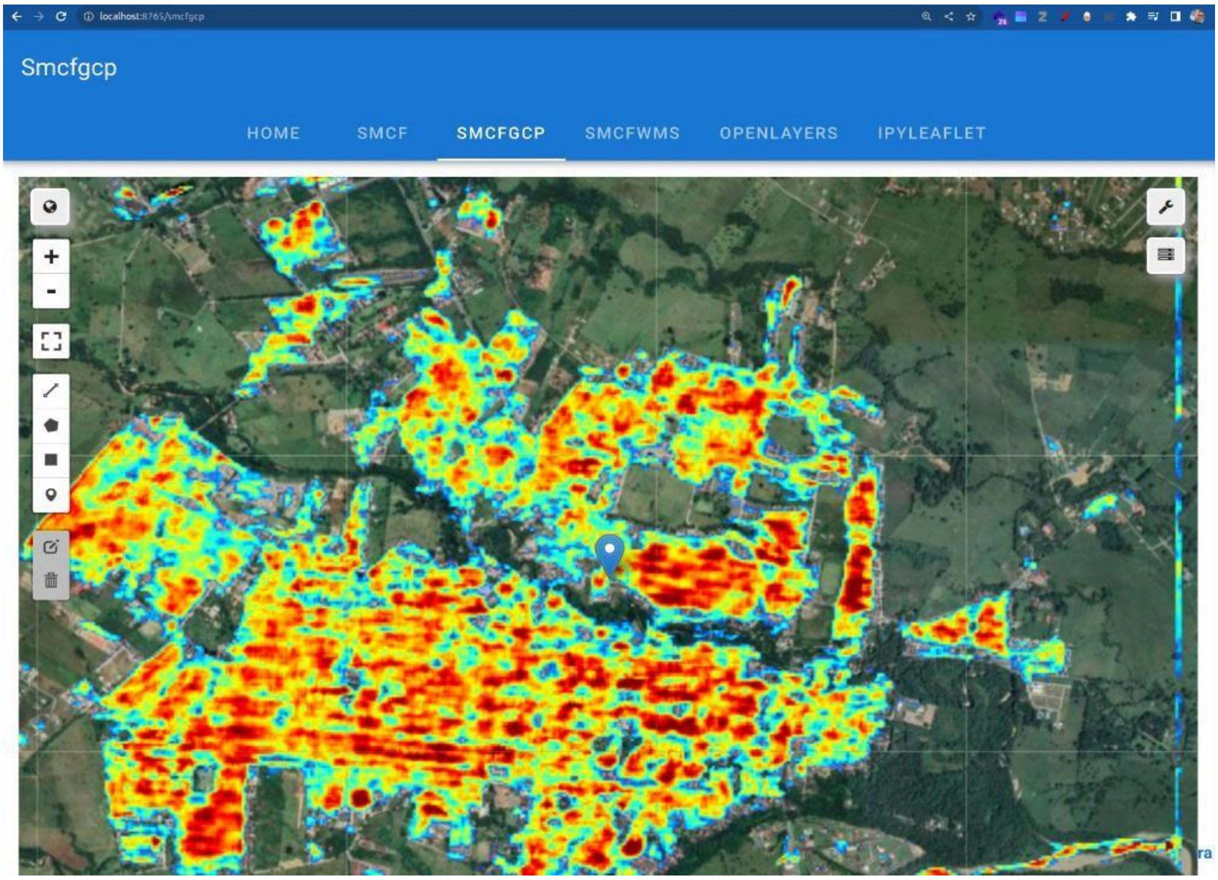

Figure 6. Visual results of urban zone reconnaissance.

The first stage of the system focuses on geospatial data processing and model building. This includes the collection and analysis of satellite imagery, census data, and other relevant sources. We use advanced remote sensing and machine learning techniques to process this data, culminating in the construction of an HDF5 file of the TensorFlow model. This file represents the synthesis of our data and analysis and is critical for the subsequent estimation and visualization stages.

The second phase of the system involves deploying the model and visualizing the results. We use Python notebooks to run the model, perform urban sprawl estimates and generate GeoTIFF files for each projected municipality. These files are published via geo services in the GeoServer application.

Finally, for interactive visualization of the results, we employed the Solara framework and Geemaps in a web application. This platform allows users to explore urban changes dynamically and in detail, offering a valuable tool for urban planning and strategic decision making (see Figure 7).

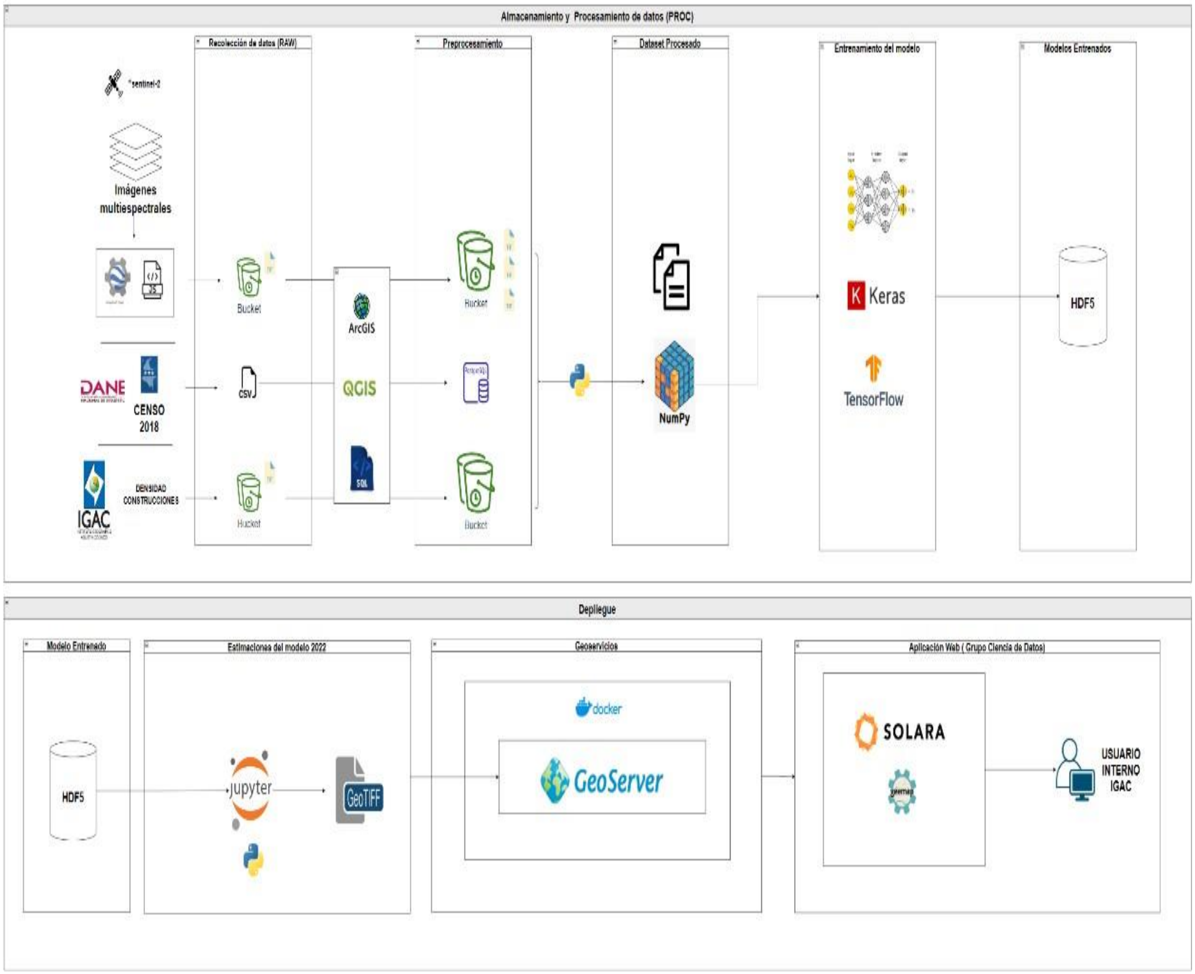

Figure 7. Architecture of the physical change monitoring system.

CONCLUSIONS:

Even with satellite images that do not have the same spatial resolution as those available through conventional providers such as Google Earth or Bing, it is evident that the Sentinel project inputs offer sufficient resolution for the model to detect significant variations in areas with urban growth. It is also important to clarify that the comparison images are a collection of different images taken during the year, with some cloud filtering criteria.

Some changes in the image may be due to several factors and not necessarily to a change in the urban growth indicator, variations in reflectance due to canopy changes, or weather conditions, but there is evidence of a detected behavior that with additional training of the model could give better results. The results shown in the application and analysis were performed with a model trained with information from 6 municipalities in the department of Meta. It was then trained with 49 prioritized municipalities, but the degree of variation did not give satisfactory results.

- **Future research:** Despite the significant achievements in our study of urban sprawl in the department of Meta, Colombia, there are several areas that could benefit from future improvements and expansions. These considerations are crucial to maintain the relevance and accuracy of the model as technologies and urban environments evolve.
- **Model Training Improvements:** To improve the accuracy and generalizability of the model, the use of more advanced training techniques should be explored. This could include implementing more sophisticated algorithms and experimenting with different hyper parameter settings.
- **Working with more Complex Data Sets:** As the study expands to more municipalities and different geographic areas, we will face the challenge of working with larger and more complex data sets. This will require improvements in processing capacity and data management techniques to ensure efficiency and accuracy in the analysis.
- **Integration of New Data Sources:** The integration of new data sources, such as advances in remote sensing and more detailed geospatial data, could significantly enrich the model. This would allow for a deeper understanding of urban sprawl patterns and their environmental and social impacts.
- **Technical and Scalability Challenges**: As the project grows, technical and scalability challenges will become more prominent. It is essential to develop solutions that can efficiently handle an increasing volume of data and ensure the scalability of the model and the technological infrastructure.

REFERENCES:

Aspinall, R. (2004). Modelling land use change with generalized linear models - A multi-model analysis of change between 1860 and 2000 in Gallatin Valley, Montana. doi:10.1016/j.jenvman.2004.02.009

Congreso de Colombia. (2019, Mayo 25). *Función Pública.* Retrieved from Ley 1955: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=93970

Congreso de Colombia. (2023, Mayo 19). *DNP.* Retrieved from Ley 2294: https://colaboracion.dnp.gov.co/CDT/portalDNP/PND-2023/Ley_2294_del_19_de_mayo_de_2023.pdf

CONPES. (2016, Junio 13). *DNP.* Retrieved from CONPES 3859: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3859.pdf

CONPES. (2019, Marzo 26). *DNP.* Retrieved from CONPES 3958: https://colaboracion.dnp.gov.co/CDT/Conpes/Econ%C3%B3micos/3958.pdf

DANE. (2020, Febrero 04). *DAPRE.* Retrieved from Decreto 148: https://dapre.presidencia.gov.co/normativa/normativa/DECRETO%20148%20DEL%2004%20DE%20FEBRERO%20DE%202020.pdf

Sui, D. (2004). Tobler's first law of geography: A big idea for a small world? doi:10.1111/j.1467-8306.2004.09402003.x

Zhang, K; Xu, G; Han, Z. (2020). Data augmentation for motor imagery signal classification based on a hybrid neural network. doi:10.3390/s20164485

BIOGRAPHICAL NOTES:

Mr. Rodolfo Yesid Meza Patacón:
Electronic engineer, specialist in project management, master in strategy and foresight. He is currently a specialized professional at the research center of the Institute of Geography Agustín Codazzi (IGAC) in Colombia, implementing Machine Learning and data science solutions.

Ms. Sonia Constanza Garzón Martínez:
Cadastral and geodetic engineer, specialist in geographic information systems, master's in information and communication sciences. She is currently a specialized professional at the research center of the Institute of Geography Agustín Codazzi (IGAC) in Colombia, leading research projects in the line of prospective analysis and data science.

CONTACTS:

Mr. Rodolfo Yesid Meza Patacón
Institute of Geography Agustín Codazzi (IGAC)
Carrera 30 # 48-51
Bogotá DC., Colombia

(+57) 3102877375
Email: rodolfo.meza@igac.gov.co
Web site: https://www.igac.gov.co/

Ms. Sonia Constanza Garzón Martínez
Institute of Geography Agustín Codazzi (IGAC)
Carrera 30 # 48-51
Bogotá DC., Colombia
(+57) 3104153520
Email: sonia.garzon@igac.gov.co
Web site: https://www.igac.gov.co/