Research practice II

Final report

A nowcasting model for Medellín city

María Camila Vásquez Correa[†], Juan Carlos Duque [‡] and Jairo Alejando Gómez[‡]
mvasqu49@eafit.edu.co, jduquec1@eafit.edu.co, jagomeze@eafit.edu.co

[†]Mathematical Engineering, Universidad EAFIT
[‡] Research group in Spatial Economics (RiSE), Universidad EAFIT

June 11, 2019

Abstract

In this article we present the first steps of a nowcasting methodology for the Medellín unemployment rate. We use two approaches, the first one include a time-series regression to identify the generating process of the unemployment series, while in the second approach we want to validate an hypothesis that correlates unemployment with the number of cars on the street. For the second approach we use an object detector to identify several categories of automobiles in images. For the first approach, we evaluated a SARIMA model to forecast the unemployment rate using historical data, and for the second approach we trained a Single-Shot Detector to detect 8 classes of vehicles in 80 CCTV traffic cameras installed in Medellín.

Keywords: Nowcasting, Unemployment, Object detection, Time series.

1 Introduction

Macroeconomic variables such as unemployment rates and GPD are key for to drive monetary policies, macro-prudential policies and fiscal policies. The real gross domestic product (GPD) is a summary of the health of an economy (Bragoli & Modugno, 2017) and is one of the most heavily monitored indicators. Also, unemployment rate, released on a monthly basis (DANE, 2018), helps the government tracking the state of the economy and, in the Colombian case, is the input for new policies and the redesign of existent ones with the aim of reducing unemployment and improving the quality of life of Colombians.

The period used to update unemployment rate and GDP does not allow the policy makers to take actions based on fresh data: they have to wait a long time until the indicator is released to make decisions with it. Our world is changing faster each day and this requires decisions with the same velocity. This implies that the conventional methodologies, such as surveys and the traditional data collection that take two months to generate an indicator as important as the unemplotment rate or the GDP, begin to be inappropriate to support the decision making.

To overcome this problem, some authors have proposed nowcasting models to estimate economic indicators with a higher frequency, allowing their respective economies to react faster. Nowcasting is a contraction of the words *now* and *forecasting* and it is defined as "the prediction of the present, the very near future and the very recent past" (Elliott & Timmermann, 2013). These models have been applied to estimate the GDP in most of the literature.

In this work, we aim to investigate a new way of producing an indicator for the unemployment rate in Medellín city, by exploring the potential of recent developments in the field of artificial intelligence, especially in convolutional neural networks, for the design and implementation of a nowcasting model that allows the appropriate authorities to measure in real time (or at least with a much higher frequency that is currently available) the unemployment rate, a key economic variable.

The underlying hypothesis is that the unemployment rate and the traffic flow on the streets are negatively correlated. In particular, the circulation of cars, buses, motorcycles, taxis and others can be an indicator of the unemployment rate. Using the object vehicle counts, along with historical data provided by key actors from the city, we aim to construct a methodology that provides a timely forecasting for the Medellín's unemployment rate.

2 State of the art

The nowcasting problem is not new. It has been treated by several authors in several countries in order to generate a real-time indicator for their decision makers, using diverse types of data and models to perform the forecast. Some of these studies are discussed in the following paragraphs.

In Canada, Bragoli & Modugno (2017) propose a model for the nowcasting of the GDP considering some variables from the United States. The problem is clear: the value of the gross domestic product of the current quarter is generated with a delay of two months, so the country has a bigger delay compared to other developed countries such as Japan or the United Kingdom. The target variable is the quarterly GDP and input series are the purchase management index, employment, manufacturing shipments, retail sales, exports and imports, among others. Many of these variables are obtained from surveys and trade indicators. In this case, the nowcasting model proposed showed better results than the forecasting made by institutional forecasters (such as the Bank of Canada or The Organization for Economic Co-operation and Development (OECD)). Also, with the variables from the United states included, the model improves drastically.

In Urasawa (2014), authors propose another model for the nowcasting of the real GDP of Japan, selecting monthly indicators such as industrial production, employment and retail sales. They follow the lead of some other authors that have models for the short-term forecast of the real GDP in various economies. For instance, they mention Lahiri & Monokroussos (2013), where authors study the effects of survey data in nowcasting the U.S GDP, and Karim *et al.* (2010), who introduce a model for nowcasting the French GDP. The conclusion in this case shows that high-frequency, real-time GDP forecasts, using a small-scale dynamic factor model, are fairly reliable for early assessment of ongoing economic activities in Japan.

It has been found that, in terms of nowcasting for macroeconomic variables, the models

proposed only include GDP. Some other examples include models for the Turkish economy, which is proposed by Modugno *et al.* (2016), and considers financial data and survey data for the nowcasting; The Czech model, detailed in Rusnák (2016) and the euro-zone model detailed in Maximo & Gabriel (2008). In all of these cases, the results outperform the forecasting made by institutional forecasters, and the models are more useful to each countries, in order to asses the current state of the economy.

For the models discussed above, the data used for the nowcasting has economic nature (surveys and financial indicators), and in most models, the measured variable is GDP. There is a great opportunity for this research project to explore ways of including new types of data (processed images from the city) and to estimate new variables (unemployment rate, for instance).

The final input for our nowcasting model will be a set of vehicle counts. However, in order to get these counts, we need to identify in a prior stage the vehicles that appear in a set of images or in a video sequence, a process known in computer vision as an object detection. A few works that approach this problem in the literature are discussed in the following paragraphs.

In (Sermanet *et al.* , 2013), a framework for using Convolutional Neural Networks (CNNs) for classification, localization and detection is presented, the presented approach is useful in order to identify multiple objects in an image. Following that work, (Girshick *et al.* , 2013) presented the Regional Convolutional Neural Network (R-CNN) framework, that aims to apply high-capacity CNNs to bottom-up region proposals in order to localize and segment objects, and then classify them into their own categories, outperforming the previous framework.

Since its introduction, the R-CNN framework has been refined. Some improvements include the Fast R-CNN, presented in (Girshick, 2015), which in comparison with Spatial Pyramid Pooling network (SPPnet) (He *et al.* , 2014) has a high performance in detection of objects, being more accurate and faster. The Fast R-CNN outperforms the R-CNN by combining the SPPnet to speed up the test time.

However, the previous approaches required a fair amount of time for computing the region proposals. Aiming to tackle this problem, (Ren *et al.* , 2015) presented an algorithmic change, based on deep CNNs to find the regions, in order to reduce the cost of the previous R-CNN approaches. This approach, which is called a Region Proposal Network (RPN), takes advantage on the high convolutional layers present in the detector as input and outputs a set of rectangular object proposals, each with an objectness score. This model is called the Faster R-CNN, and it was extended by the Mask R-CNN (He *et al.* , 2017).

Using a completely different approach, (Liu *et al.* , 2015) introduced the Single-Shot detector (SSD), which outperforms the previous releases with a simpler model, easier to train and straightforward to integrate with other systems. It is one of the favorite approaches for real time processing of images, due to its test time and high accuracy.

One of the most interesting applications for object detection is the detection and tracking of objects in video. Some approaches to perform this task are the You Only Look Once (YOLO) architecture (Shafiee *et al.* , 2017), available in three versions, as well as a SqueezeNet architecture

(Wu *et al.* , 2016), that perform well in real-time object detection. Table 1 shows the comparison of each one of these models.

Table 1: Comparison of object detection Models. PASCAL VOC dataset.

Model	**Mean Average Precision (mAP)**	**Advantages**	**Disadvantages**
SSD	80.0	Has troubles dealing with small data sets.	Low inference time: suitable for real time inference.
Faster-RCNN	75.9	Great accuracy in small data sets.	High inference time.
Yolo	57.9	Low general accuracy.	Great performance in real time inference.
SqueezeNet	70.5	Small network, fewer parameters: less training time. Also, less inference time.	Large data sets required.

3 Theoretical Framework

3.1 Unemployment rate characterization

The unemployment rate is an indicator of the degree of use of human resources in the economy (DANE, 2016), it is also an indicator of the life quality of the population and its performance in society. For this reason, the unemployment rate is an important indicator to take into account in the design of public policies for Colombian cities. This rate is released by the DANE in a monthly basis, based on a mobile quarter measurement, with a two-month delay.

Going trough the methodology of the National Administrative Department of Statistics, DANE, we found the following information about the calculation of the unemployment rate in the country for the series reported for the years 2000 trough 2018 (DANE, 2016). This department performs a continuous survey in the country, called the "Large integrated household survey" (Gran Encuesta Integrada de Hogares), or GEIH, for its Spanish acronym, which provides the unemployment rate among other important information on a monthly basis.

3.1.1 Calculation of the unemployment rate

The unemployment rate, UR, is the ratio between the number of people seeking for a job and the number of people in the labor force:

$$UR = \frac{DS}{EAP} * 100 \tag{1}$$

Where EAP stands for Economically Active Population and DS is the number of people seeking for a job. This last portion of the population is extracted from the GEIH. It is composed by two

additional variables:

- *Open unemployment rate (OUR)*:

$$OUR = \frac{DSA}{EAP} * 100 \tag{2}$$

  DSA stands for the number of people that were:

  - Unemployed in the reference week.
  - Did errands in the last month.
  - Available.

- *Hidden unemployment rate (HUR)*

$$HUR = \frac{DSO}{EAP} * 100 \tag{3}$$

  DSO stands for the people that were:

  - Unemployed in the reference week.
  - Did not do errands in the last month, but in the last 12 months.
  - Available.

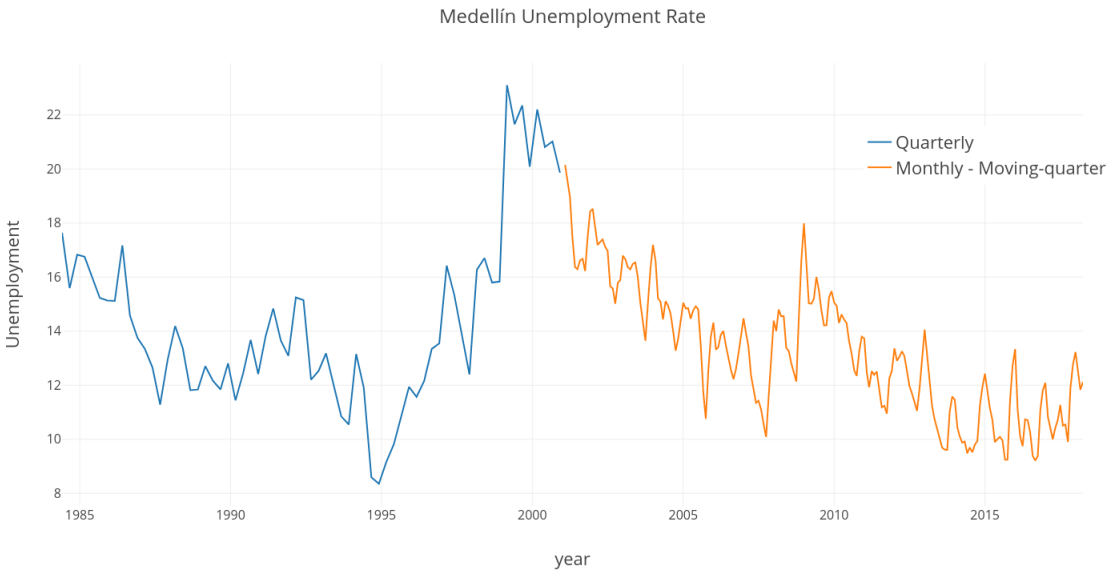


Figure 1: Time series of the unemployment rate of Medellín and its metropolitan area.

For Medellín, which is one of Colombia's capital cities, the unemployment rate is calculated every month for the city and its Metropolitan Area. The collection period is weekly, and is designed in a way that allows the department to collect the necessary data for a good forecast of the unemployment rate, while reducing time and cost. The criteria used for choosing the households to be surveyed guarantees, based on statistics, a relative standard error of less than 5%, and also the

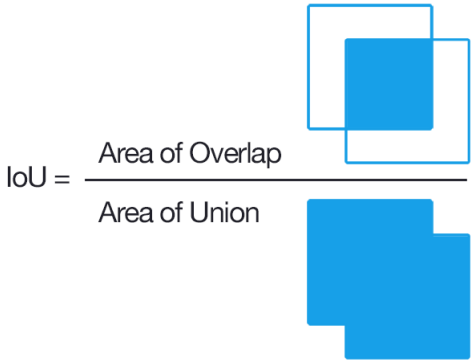


Figure 2: Computing the Intersection of Union is as simple as dividing the area of overlap between the bounding boxes by the area of union. Source: Rosebrock (2016).

sample design allows segments to not be repeated in the capital in three years, ensuring sufficient variability in the sample to achieve reliable results (DANE, 2016).

This survey was updated in 2000. Before that year, the unemployment rate was calculated using another survey, and it provided the data quarterly, not in a moving quarter. The left graph of Figure 1 shows the historical data for the unemployment rate from 1994 to 2000, in orange, and in blue the one from 2000 to August 2018. Due to the sample period difference, we are taking the data from 2000 for analysis purposes.

3.2 Object detection metrics

The performance of an object detection model can be measured using metrics such as the mean average precision (mAP) and the Precision $\times$ Recall curve. To define this metrics, we must introduce the following concepts:

- **Intersection Over Union (IOU):** Is a measure based on Jaccard Index (Real & Vargas, 1996) that evaluates the overlap between two bounding boxes, just as shown in Figure 2. It requires a ground truth bounding box B_{gt} and a predicted bounding box B_p. By applying the IOU we can tell if a detection is valid (True Positive) or not (False Positive):

$$IOU = \frac{\text{area}\,(B_p \bigcap B_{gt})}{\text{area}\,(B_p \bigcup B_{gt})} \tag{4}$$

- **True Positive (TP):** A correct detection. A detection with IOU$\geq$threshold.
- **False Positive (FP):** A wrong detection. A detection with IOU$<$threshold.
- **False Negative(TN):** A ground truth bounding box that was not detected.

With these concepts, is possible to define the *precision* and *recall* as follows:

$$Precision = \frac{TP}{TP + FP} \qquad\qquad Recall = \frac{TP}{TP + FN} \tag{5}$$

The Precision×Recall curve is used to measure the performance of the model. An object detector of a particular class is considered good if its precision stays high as its recall increases, which means that if you vary the confidence threshold, the precision and recall will still be high. Another way to identify a good object detector is to look for a detector that can identify only relevant objects (0 False Positives = high precision), finding all ground truth objects (0 False Negatives = high recall) (Davis & Goadrich, 2006).

Another metric for the detector is the **Average precision**(AP), which is the area under the Precision×Recall curve. This metric gives an insight on how well the model is behaving, in general terms, for each class. The **mean average precision**(mAP) does the same for all the classes in the training data set, by averaging the AP for each class. The State of the Art for this metric in object detection models is currently in 48.4% for the COCO dataset.

3.3 Multivariate time series models

3.3.1 Cointegration

Cointegration is a statistical property of a collection $(X_1, X_2, \ldots, X_k)$ of time series variables. If (X, Y, Z) are all integrated of order d, and there exists some constants a, b, c such that $aX + bY + cZ$ is integrated of order $d_1 < d$, then the series X, Y and Z are co-integrated (Granger, 1981). If two or more series are cointegrated, this cointegration can be a test for the statistical significance of a connection between this series (Granger & Newbold, 1974). A non cointegration between two series can lead to spurious correlation, which we would like to avoid (Yule, 1926).

Some cointegration tests include:

- *Engle–Granger two-step method*: If X_t and Y_t are non-stationary and integrated of order 1, we can find the linear combination of them both that is stationary via Ordinary Least Square estimation. In other words, with the linear combination:

$$Y_t - \beta X_t = u_t \tag{6}$$

  Where u_t is stationary, we want to estimate β. Then, we can test stationarity for u_t, and conclude wether or not X_t and Y_t are cointegrated (Engle & Granger, 1987).

- *Johansen test:* This test allows to test cointegration in more than 2 series. It also permits mote than one cointegrating relationship between the series (Johansen, 1991).

3.4 Vector autoregression models

A vector autogregression model (VAR) describes the evolution of a set of k variables over the same sample period as a linear function of their past values. These variables are contained in a vector of k dimensions Y_t, where the i^{th} element is the observation at time t of the i^{th} variable. A $p - th$ order VAR (VAR(p)) is defined as follows (Wei, 2006):

$$Y_t = c + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots A_p Y_{t-p} + \epsilon_t \tag{7}$$

Where:

- c is a k-vector of constants.
- A_i is a $(k \times k)$-matrix of constants.
- ϵ_t is a k-vector of error terms with mean 0 and no correlation.

For this model, all variables must be the same order of integration. If all of them are stationary or integrated of order $d > 0$, they must not be cointegrated. In other case, the model must be restricted and should include the error correction term. The model then turns into a Vector Error Correction model (VEC) (Engle & Granger, 1987).

3.5 Recurrent Neural Networks

A recurrent neural network (RNN) is a class of artificial neural network where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. This networks can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition or time series forecasting (Miljanovic, 2012).

4 Methodology

This section describes in detail the steps needed to build a nowcasting model using traffic from Medellín city. Figures 3 and 4 describe the workflow for this research, which is explained in this section. The project is developed in 3 main stages, during one year and a half.

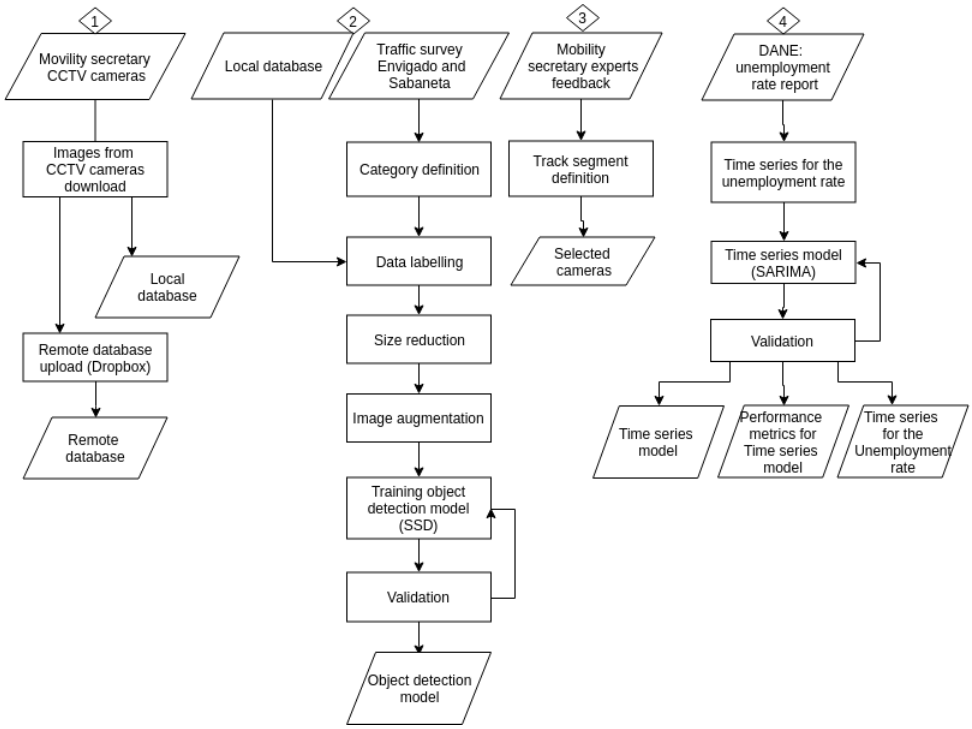


Figure 3: Main steps for forecasting the unemployment using traffic, Stage 1.

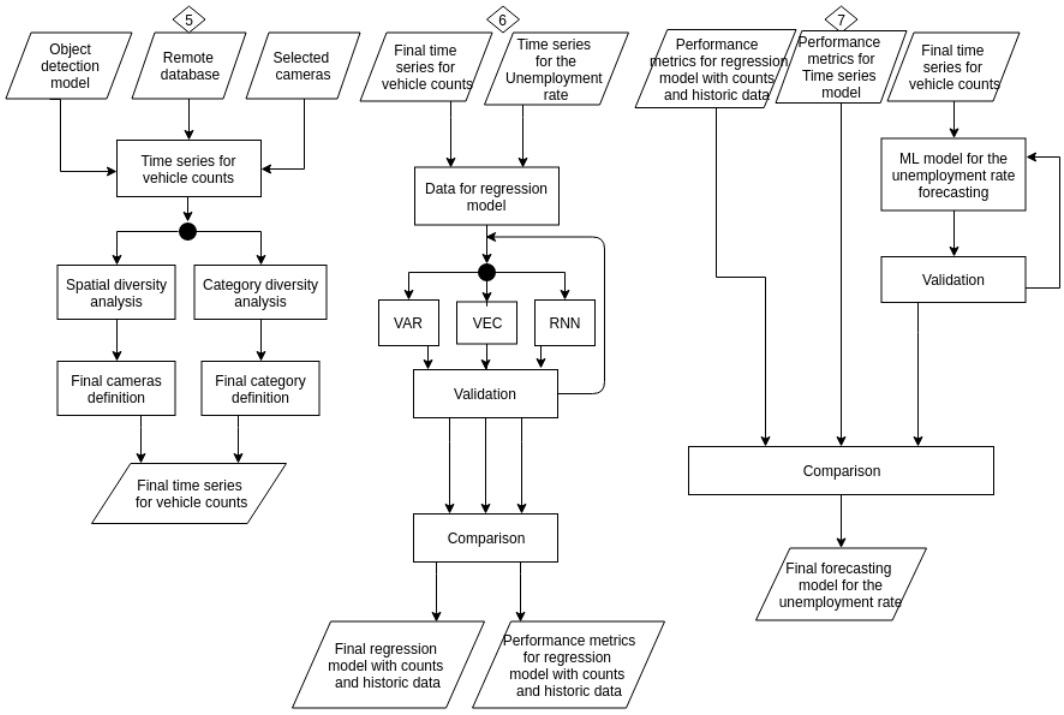


Figure 4: Main steps for forecasting the unemployment using traffic, Stage 2.

4.1 Data

The step 1 in Figure 3 corresponds to the data collection for the model. First, we scraped images from the CCTV cameras of Medellín. A total of 80 cameras are available from the mobility secretary website (secretaría de movilidad de Medellín, 2018). These images were stored in a local database, and later uploaded to Dropbox, in order to build a remote database for further use. These images are downloaded every 2 minutes, corresponding to the refresh period of the cameras on the Web site.

Figures 5 and 6 show snapshots of some of the cameras at different times. First, we note that, at times, some cameras are not available, or they get frozen at a certain time, delivering the same image during multiple time stamps. Then, some cameras have lower resolution than others, and the traffic changes along the moment of the day and the section we are looking at. A final annotation is that some cameras have unuseful angles, such as the one in the left top corner in Figure 6. All of these are issues to consider when producing vehicle count estimates from the cameras.

To try and fix some of these issues, we selected some strategic streets from the city, that provide the most information about the current activity. The selection was made based on the recommendations made by the Mobility Secretary of Medellín and one external expert in Mobility, and correspond to the third step in Figure 3. The selected road tracks include:

1. San Juan Avenue: Strategic relevance due to the high flow of buses, and its proximity to a commercial point in the city, which generates high-traffic flow from industrial sectors, wood, leather, metalworking clusters and its proximity to the utility company of Medellín (EPM).

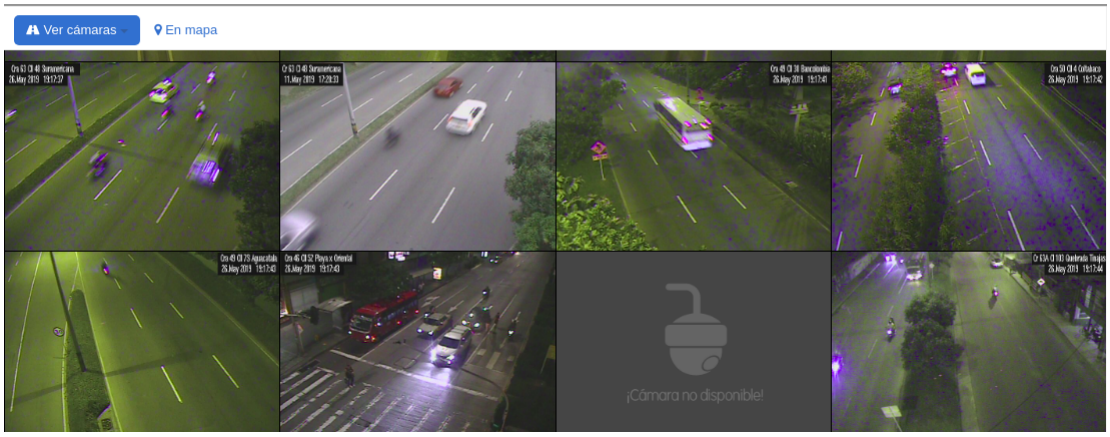


Figure 5: Sample images taken from the cameras located in Medellín, Night.

Figure 6: Sample images taken from the cameras located in Medellín, Day.

2. North freeway: Relevant for its connection with the Colombian coast and the capital of the country.
3. South freeway: Relevant for its connection with other municipalities in the metropolitan area and the South of the country.
4. Iguaná: It routes part of the vehicular flow that comes from the West.
5. 80th Avenue: Located on the Western side of the city. Captures the flow of vehicles from North to South.
6. 33 Avenue: Located on the Western side of the city. Captures the flow of vehicles from East to West.

These sectors contribute to analyze the economic flow of the city, the inputs and outputs from

other regions of the country and the traffic. However, we also included other important corridors to analyze the intra-urban road traffic. These segments reflect the internal vehicular traffic, and they could be useful to capture the commerce of the city itself and people mobility for job-related reasons:

- Las Palmas Avenue.
- Avenue El Poblado.
- 30th Avenue.
- Avenue Guayabal.

With the data regarding the unemployment rate, we obtained the official rate reported by the national department of statistics (DANE, for its Spanish acronym) (DANE, 2018). This data set is described in further sections.

4.2 Time series analysis

In order to characterize the unemployment rate, the first step was to perform a complete time-series analysis aiming to understand its behavior and quantify the dependence of the series from past observations. This approach will lead us to to design a basic prediction model that could serve as a baseline to measure our progress in the nowcasting process, once we add the vehicle counts. Also, it could give us an idea of the behaviors that the counts series should capture, or how the counts should be taken to fully reflect the characteristics of unemployment of the city.

Following the proposed methodology in (Wei, 2006), we are searching certain patterns in the series that allow us to identify whether or not they are stationary, if they are auto correlated, have moving average, present seasonal pattern and if they have an stable long-term variance or they present heteroscedasticity. After identifying these characteristics, we will be able to find a possible generating process of the series, and propose a first forecasting model for it, following step 4 in Figure 3.

4.3 Object Detection Model

For the cameras situated in strategic places of the city, our objective is to identify the number of vehicles of different types circulating the city in various periods of time. For this purpose, we selected some important information to be extracted from the images of the cameras of the city: the type of vehicles we want to detect, count and generate a time-series of counts with. The categories of interest are:

- Motorcycles of any type, bicycles and three wheeler.
- Car, all terrain vehicle (ATV).
- Van.
- Taxi.
- Camper.
- Truck, tractor, dump truck.
- Bus, minibus.
- Industrial machinery.

These categories were defined by exploring the existing and registered automobiles in the cities of Envigado, Copacabana and Sabaneta, which are part of the Metropolitan area of Medellín. The data can be found in (Datos Abiertos, 2018a), (Datos Abiertos, 2017) and (Datos Abiertos,

2018b). The *industrial machinery* category is formed by the following subcategories (appearing in (Datos Abiertos, 2018a)):

- Froklitft.
- Farm equipment.
- Backhoe.
- Tractor.
- Excavator.
- Charger.
- Front loader.
- Semitrailer.

The ambulance category is discarded due to the fact that no economic information is provided by it, and it will be considered within the *van* or *car* category, according to the case. Also, we merged some categories that could be considered small, such as bicycles and three Wheeler with motorcycles, for analysis purposes.

With this categories defined, we then proceed to label images from our local database, to build a training set and re-training an object detection model. In Figure 7 there are some examples of labelled images. There are in total 2000 images in the data set, in two different moments of the day and taking images from all the possible cameras.

(a) (b)

Figure 7: Some labelled images from the local database.

The model we chose was a Single Shot Detector (SSD) over a Faster Regional Convolutional Neural Network (RNN), due to the fact that it has a better performance in terms of test time, and it has a good accuracy. Also, there are some SSD architectures already trained to identify characteristics from vehicles, and detect categories such as cars, buses and motorcycles in images that served as a first prototype approach to perform the counts of the vehicles circulating in Medellín.

While SSD are designed for object detection in real-time, Faster R-CNN uses a region proposal network to create boundary boxes and utilizes those boxes to classify objects. Although it is considered the start-of-the-art in accuracy, the whole process runs at 7 frames per second. Far below what a real-time processing needs. SSD speeds up the process by eliminating the need of

the region proposal network. To recover the drop in accuracy, SSD applies a few improvements including multi-scale features and default boxes. These improvements allow SSD to match the Faster R-CNN's accuracy using lower resolution images, which fimproves the speed further. According to the following comparison, it achieves the real-time processing speed and even beats the accuracy of the Faster R-CNN (Liu *et al.* , 2015).

In order to make use of the available Single Shot Detectors, a Transfer Learning (Pan & Yang, 2010) approach is adopted. It makes use of previously learned features by the network to help it identify new categories. To perform this step, it is necessary to consider the new categories of interest, build a data set that contains a representative number of examples per class, and follow the transfer learning process, that takes the training of the network in a certain step, called a checkpoint, and introduces the new training and test data.

Along with the training, some steps were performed in the data set so that it could train a better detection model. Firstly, we reduced the size of the images, to reduce the training time, to a size where the categories were still recognizable, so it would not affect the predictions. Then, we perform an image augmentation process. This process applies a set of transformations to the original images in the training set. The transformations applied were:

- Rotation: it captures the different angles in images.
- Sharpen and emboss: can capture some of the resolution variations.
- Crop and pad: helps with more examples of vehicles appearing in corners, or cropped in the camera.
- Dropout (randomly deleting pixels in an image): this makes a new image that is completely different for the Network, but is essentially the same and contains the same information.
- Motion Blur: this was used to emulate the instances where vehicles go at a high speed.
- Color change: this can capture the changes in lighting in the images.

Once we applied the data augmentation, we ended up with a 400,000 image data set that we use to train a SSD, starting from a specific checkpoint, in Apolo Scientific Computing Center, at EAFIT University in Medellín. Firstly, we tuned the hiper-parameters of the model, such as the learning rate, the batch size and the maximum number of steps. Finally, with the chosen parameters, we perform a complete training.

4.3.1 Validation

In order to evaluate the performance of the detector, we use precision and recall. Precision is the ability of a model to identify **only** the relevant objects, while recall is the ability to find all the relevant cases (Manning *et al.* , 2010). After this validation process, we will conclude the second step in Figure 3 and will be able to produce vehicle counts for each defined category. Counting these vehicles will allow us to have a time series that provide the information about the automobiles in the city with different sample periods (dayly, weekly and monthly), that will finally be the input for the conjoint regression model in step 6 and will be analyzed in step 5 in Figure 4.

4.4 Conjoint regression models

Using the series for the counts of vehicles obtained with the object detection model from previous section, by performing an inference of the vehicles in each image in the selected cameras from the remote database, the objective of this section is to describe: a way to analyze these series behavior, just as step 5 in Figure 4 and a regression model that forecasts the unemployment rate using the historical data and the counts of vehicles, as shown in step 6 of Figure 4.

4.4.1 Data analysis

The pre-processing of the time series for the counts of vehicles will include tests to see whether they are: cointegrated (if their regression is statistically significant) or redundant (if they provide the same information). The spatial and categorical diversity will be tested and the redundant series will be discarded. These analysis will be done trough the cointegration tests, to define how can we use these series to forecast the unemployment.

This analysis will define the final time series for vehicle counts, that will serve as input, along with the time series for the unemployment rate, for the regression models we describe below.

4.4.2 Proposed models

A Vector Autorregression model (VAR) and a Vector Error Correction model (VEC) will be evaluated with the series for the counts of vehicles and the historical unemployment rate, once the collection and preprocessing is done. This will allow us to compare the results from previous section and determine the next step. The VAR model will be useful when the series are no cointegrated, and in that case it will show whether the series can explain the unemployment or if they can be discarded. In the cointegration case, the VEC model will be useful in order to determine if the series are correlated and can help us explain the unemployment. Also, in the next stage of the research, we will evaluate a recurrent neural network model to perform the regression, via training with the available information: the unemployment rate time series and the series from vehicle counts.

The validation of this models will compare their results with the official report of the unemployment rate, to see how well the model performs and if it can reach somehow the reports. This comparison can be done using the Mean Squared error for the predictions, as it was done for the time series model. Finally, with the validated models, one will be chosen by comparing their performance metrics. This will be the final regression model with vehicle counts and historical data, and this concludes the step 6 in Figure 4.

4.4.3 Prediction with counts

The final stage of this research, as shown in step 7 of Figure 4, consists in the search for a method that will forecast the unemployment rate using only the counts of traffic in the city. This will be a Machine Learning model soon to be defined, or an economic model. This model will be validated and compared to the performance of the time series model and the regression model with counts and historic data, with the purpose of determining the best models among all the constructed ones.

5 Results

5.1 Time Series Analysis

Looking at the series, in Figure 1, and evidencing its clear trend, the first approach is to differentiate the series and observe that the differentiated series are stationary, while the original are not. Figure 8 illustrates the autocorrelation and partial autocorrelation graphs of the series, which show evidence of auto-regressive component and moving average, along with a clear evidence of hysteresis.

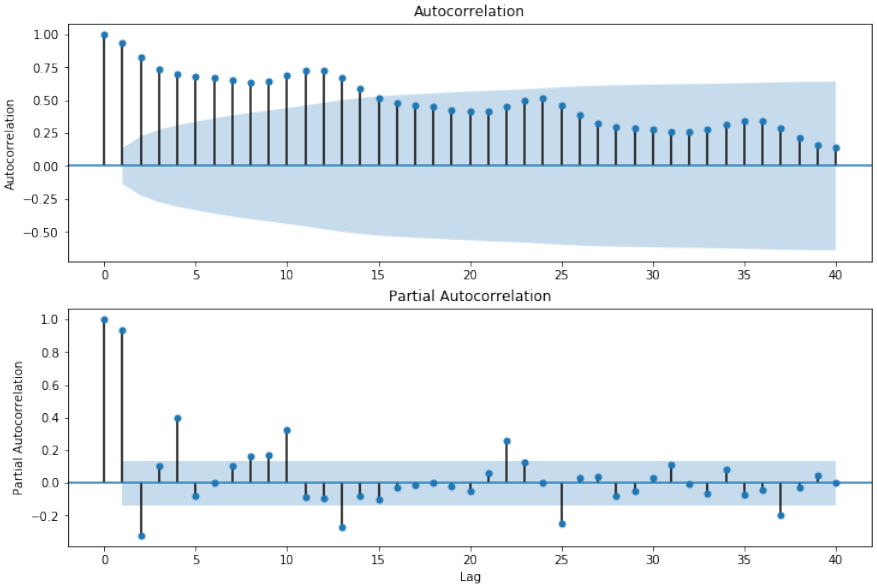


Figure 8: Autocorrealtion and partial autocorrelation graph for the series.

Hysteresis is defined as a condition in which a system state does not depend only on the external conditions at the moment considered, but also on the evolution followed until reaching said state (González, 2012). In several studies, it has been proved that the unemployment rate for Medellín city has clear signs of this phenomenon. The hysteresis of the series reflects in an absence of an effect of monetary policy and a regionally differentiated impact of investment and exports on this rate are palpable (González, 2012).

Given the behavior of the city economy across different periods, we can intuitively consider a seasonality hypothesis, in which unemployment rises or falls depending on the needs of the market in a certain period of the year. With these resources, (the non-stationarity of order one for the series, the autocorrelation and moving average component and the seasonal pattern), we proceed to fit a seasonal, auto regressive, integrated, moving average SARIMA model for the series Wei (2006). Before trying the combinations of parameters that could suit our model, the estimation results for different values of p, d, q and s are shown in Table 2. The best model for each s was chosen using also the Akaike's information criteria (AIC) (Akaike, 1974).

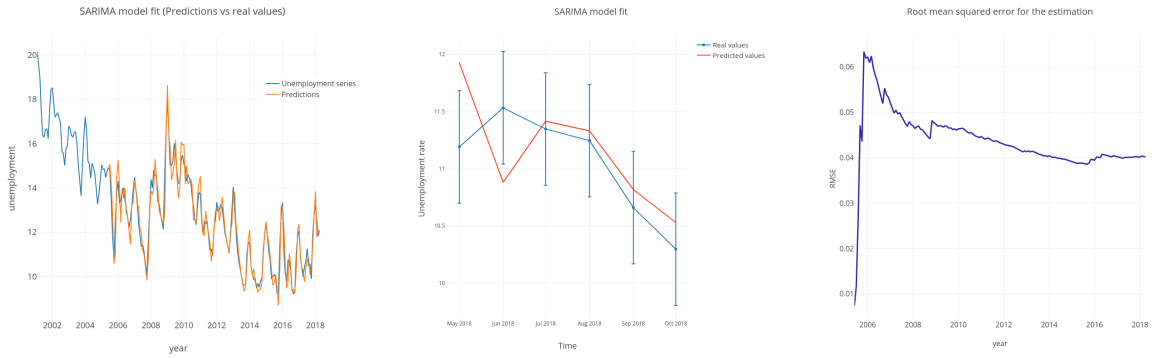
We choose the SARIMA$(2, 1, 2) \times (2, 0, 1, 12)$ model, even though, when making the regression, the second lag for the autoregressive seasonal component is only significant with a 10%. In

Table 2: Results for the model estimation.

Model	AIC
SARIMA(2,1,2)x(2,0,1,12)	-778.54
SARIMA(1,1,2)x(2,0,2,6)	-761.74
SARIMA(1,1,0)x(1,0,2,3)	-727.11
ARIMA(2,1,2)	-665.27

Figure 9a, are shown the results of fitting the model to a portion of the data an predicting the next values from it. In Figure 9b it is shown the performance of the model for new historical values. Also, Figure 9c is shows the root mean squared error for these predictions. By seeing the scale, one can notice that the error is really low, and it decreases as the prediction values are added to the training data.

From this analysis, we found some interesting characteristics of the unemployment: it depends on its previous values, as well as in the particular behavior of the cities economy, and the policies designed for its control will not be reflected in the short term. Also, the data collection and processing made by our approach should consider these trends and behaviors, and the seasonality that includes a whole year in order to capture the whole phenomenon of the unemployment rate. This is the first forecasting model proposed, following a classical approach, that we intend to improve.



(a) Predictions for the model on a rolling window.

(b) Predictions on new values.

(c) Root mean squared error for the predictions.

Figure 9: Predictions of the unemployment rate using a SARIMA$(2,1,2) \times (2,0,1,12)$ model.

5.2 Object detection

The training an validation images were labelled, a total of 2000 training images (split in training and test set) and 200 validation images. With this data set, we retrained the MobileNet SSD. In our first stage, the results were not the ones we expected, so we reduced the size of our input images by 40% and added the augmented images to the training set. The Figure 10 shows some of the image augmentations we performed for the dataset, as described in te previous section.

Figure 10: Some tranformations applied to the original training data set.

The results in terms of validation for the first training are shown in Figure 11. Here, we can see the Precision $\times$ Recall curve for each class, in which some classes are clearly performing better than others, in terms o the curve and the mAP. The car class is the best so far, probably because it is better represented in the data set. Also, the Industrial Machinery does not show any data, as the model is incapable of detecting a vehicle with this label.

With the synthetic images obtained via augmentation, the model does not perform better than the previous one. In fact, after several changes in parameters and fixing issues, the performance decayed in comparison to the first trained model. Analyzing the architecture, it could probably have learned the augmentation, but it does not show evidences of over fitting, that could be evidenced in a decaying training loss and an static or increasing test loss.

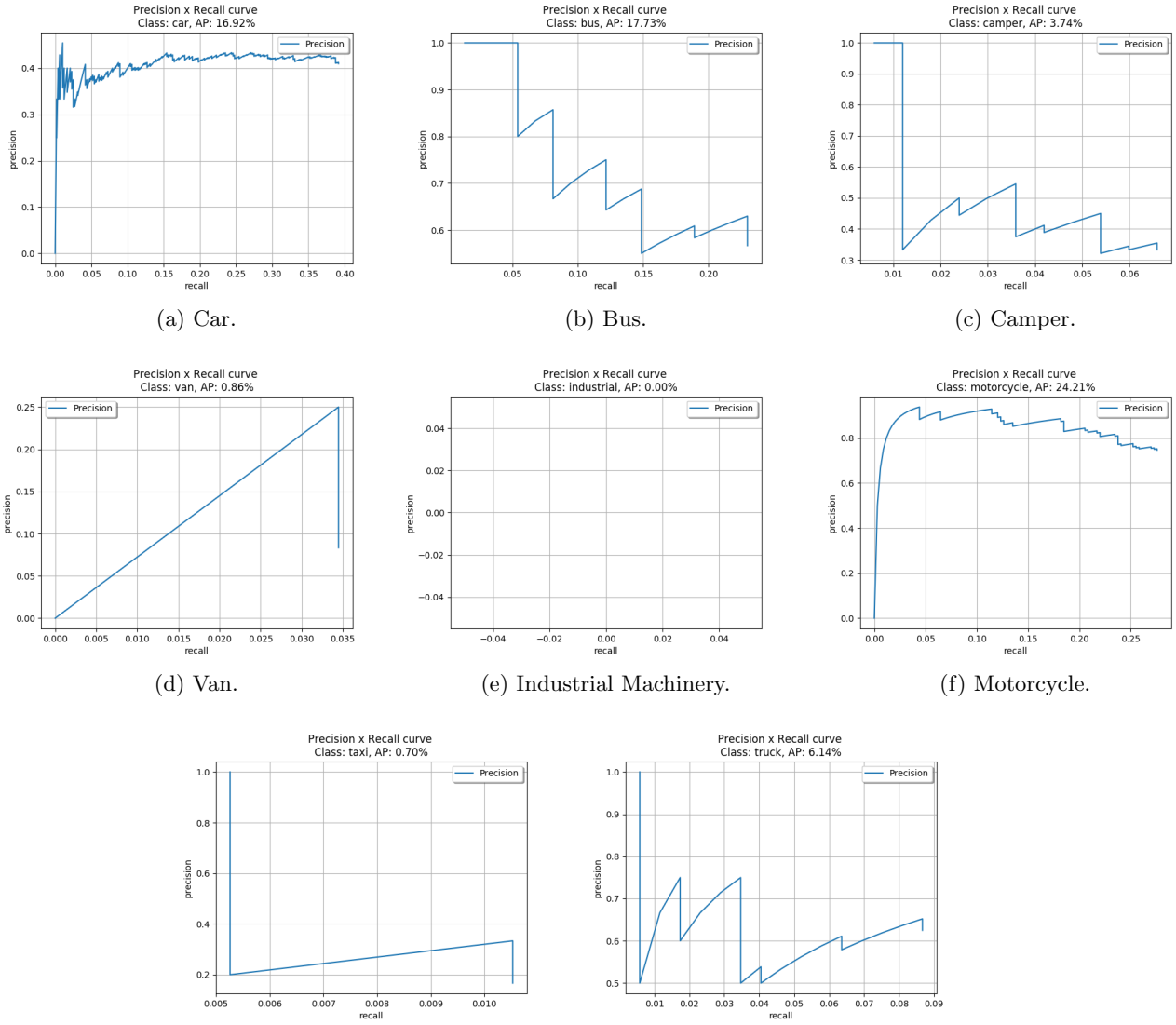
In future research, we expect to fix this issue, by labelling more images and probably trying out other architectures that could perform better in our problem. Finally, we expect to have an object detection model for the classes we defined. With this model, we will generate the time series that will be used by the new regression model for the unemployment rate. We want to see if this new information provides an improvement of the time series model. Also, we want to see if the vehicles in the city help explain the unemployment, to build our final predictor model using only counts.

6 Conclusions and future research

The series for the unemployment are complex and have dynamics that should be taken into account when predicting them with another data source, such as the hysteresis and the seasonality found in the time-series analysis. The generating process found is a good first step for the forecasting, and could be used as an input for the nowcasting by vehicle counts, by measuring the correlation between this series and the traffic in the city, and proposing a new forecasting model, via time-series or machine learning, whose inputs could include the historic unemployment rate and the series for the counts.

A Single Shot Detector was trained to obtain counts from cameras of the city. This detector will generate the necessary inputs for the conjoint regression models with counts and historical data, when the performance is improved.

For future research, we expect to generate new forecasting models for the unemployment rate that include the series for the counts of the vehicles from the city, one that gets fed up with the



(a) Car. (b) Bus. (c) Camper.

(d) Van. (e) Industrial Machinery. (f) Motorcycle.

(g) Taxi. (h) Truck.

Figure 11: Precision × recall curve for each class.

historical unemployment, and one that could predict the unemployment using only the vehicle counts, without the historical unemployment rate data. The reasons for the incremental steps are firstly, to have an insight of the generating process of the unemployment, its characteristics and the phenomenons that can affect it. Secondly, before starting to build a forecasting model with vehicle counts, we want to analyze whether or not the series of the vehicle counts and the unemployment rate are correlated and provide useful information. And finally, the goal is to generate the nowcasting model only with the counts, knowing they provide the information of the unemployment and capture its main characteristics.

Acknowledgements

To the researchers at RiSE group, who provided support for the development of this project. Also, to the computational biology group, the center of scientific computation Apolo, and Santiago Hincapié-Potes, for their help in this practice.

References

Akaike, Hirotugu. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Bragoli, Daniela, & Modugno, Michele. 2017. A now-casting model for Canada: Do U.S. variables matter? *International Journal of Forecasting*, **33**(4), 786 – 800. "http://www.sciencedirect.com/science/article/pii/S0169207017300341".

DANE, Departamento Administrativo Nacional de Estadística. 2016 (april). *Metodología General Gran Encuesta Integrada de Hogares - GEIH.* https://www.dane.gov.co/files/investigaciones/fichas/metodologia_GEIH-01_V9_2.pdf.

DANE, Departamento Administrativo Nacional de Estadística. 2018 (Jun). *Gran encuesta integrada de hogares (GEIH) Mercado laboral.* https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo.

Datos Abiertos, Gobierno digital Colombia. 2017 (Ago). *Distribucion Del Parque Automotor Inscrito2016(II).* https://www.datos.gov.co/Transporte/Distribucion-Del-Parque-Automotor-Inscrito2016-II-/nd4h-ytqu.

Datos Abiertos, Gobierno digital Colombia. 2018a (Jul). *Parque Automotor Municipio de Envigado hasta 31-07-2018.* https://www.datos.gov.co/Transporte/Parque-Automotor-Municipio-de-Envigado-hasta-31-07/fqwq-ik2w/data.

Datos Abiertos, Gobierno digital Colombia. 2018b (Oct). *Registro histórico del parque automotor de Sabaneta.* https://www.datos.gov.co/Transporte/Registro-hist-rico-del-parque-automotor-de-Sabanet/bewj-2mvn.

Davis, Jesse, & Goadrich, Mark. 2006. The relationship between Precision-Recall and ROC curves. *Pages 233–240 of: Proceedings of the 23rd international conference on Machine learning.* ACM.

Elliott, G., & Timmermann, A. 2013. *Handbook of Economic Forecasting.* Handbook of Economic Forecasting. Elsevier Science. https://books.google.com.co/books?id=-_dSuakNHWYC.

Engle, Robert F., & Granger, C. W. J. 1987. Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, **55**(2), 251–276.

Girshick, Ross B. 2015. Fast R-CNN. *CoRR*, **abs/1504.08083**. http://arxiv.org/abs/1504.08083.

Girshick, Ross B., Donahue, Jeff, Darrell, Trevor, & Malik, Jitendra. 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, **abs/1311.2524**. http://arxiv.org/abs/1311.2524.

González, Angélica Beltrán. 2012. *Histéresis en el desempleo: una revisión de los estudios para Colombia*. Universidad Colegio Mayor Nuestra Señora del Rosario. http://repository.urosario.edu.co/bitstream/handle/10336/3947/52057350-2012.pdf;jsessionid=CF10FC014D8D1CBE40F3918925EEAFAF?sequence=1.

Granger, Clive WJ. 1981. Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, **16**(1), 121–130.

Granger, Clive WJ, & Newbold, Paul. 1974. Spurious regressions in econometrics. *Journal of econometrics*, **2**(2), 111–120.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. *Pages 346–361 of: European conference on computer vision*. Springer.

He, Kaiming, Gkioxari, Georgia, Dollár, Piotr, & Girshick, Ross B. 2017. Mask R-CNN. *CoRR*, **abs/1703.06870**. http://arxiv.org/abs/1703.06870.

Johansen, Søren. 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, 1551–1580.

Karim, Barhoumi, Olivier, Darné, & Laurent, Ferrara. 2010. Are disaggregate data useful for factor analysis in forecasting French GDP? *Journal of Forecasting*, **29**(1-2), 132–144. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.1162.

Lahiri, Kajal, & Monokroussos, George. 2013. Nowcasting US GDP: The role of ISM business surveys. *International Journal of Forecasting*, **29**(4), 644 – 658. http://www.sciencedirect.com/science/article/pii/S0169207012000374.

Liu, Wei, Anguelov, Dragomir, Erhan, Dumitru, Szegedy, Christian, Reed, Scott E., Fu, Cheng-Yang, & Berg, Alexander C. 2015. SSD: Single Shot MultiBox Detector. *CoRR*, **abs/1512.02325**. http://arxiv.org/abs/1512.02325.

Manning, Christopher, Raghavan, Prabhakar, & Schütze, Hinrich. 2010. Introduction to information retrieval. *Natural Language Engineering*, **16**(1), 100–103.

Maximo, Camacho, & Gabriel, Perez-Quiros. 2008. Introducing the euro-sting: Short-term indicator of euro area growth. *Journal of Applied Econometrics*, **25**(4), 663–694. https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1174.

Miljanovic, Milos. 2012. Comparative analysis of recurrent and finite impulse response neural networks in time series prediction. *Indian Journal of Computer Science and Engineering*, 180–191.

Modugno, Michele, Soybilgen, Barış, & Yazgan, Ege. 2016. Nowcasting Turkish GDP and news decomposition. *International Journal of Forecasting*, **32**(4), 1369 – 1384. http://www.sciencedirect.com/science/article/pii/S0169207016300693.

Pan, S. J., & Yang, Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**(10), 1345–1359.

Real, Raimundo, & Vargas, J. 1996. The Probabilistic Basis of Jaccard's Index of Similarity. **45**(09), 380–385.

Ren, Shaoqing, He, Kaiming, Girshick, Ross B., & Sun, Jian. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR*, **abs/1506.01497**. http://arxiv.org/abs/1506.01497.

Rosebrock, Adrian. 2016 (November). *Intersection over Union (IoU) for object detection*. https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/.

Rusnák, Marek. 2016. Nowcasting Czech GDP in real time. *Economic Modelling*, **54**, 26 – 39. http://www.sciencedirect.com/science/article/pii/S0264999315004034.

secretaría de movilidad de Medellín. 2018. *Mapas SIMM*. https://www.medellin.gov.co/simm/mapas/index.html?map=camarasCctv. Accessed: 2018-10-03.

Sermanet, Pierre, Eigen, David, Zhang, Xiang, Mathieu, Michaël, Fergus, Rob, & LeCun, Yann. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

Shafiee, Mohammad Javad, Chywl, Brendan, Li, Francis, & Wong, Alexander. 2017. Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video. *CoRR*, **abs/1709.05943**. http://arxiv.org/abs/1709.05943.

Urasawa, Satoshi. 2014. Real-time GDP forecasting for Japan: A dynamic factor model approach. *Journal of the Japanese and International Economies*, **34**, 116 – 134. http://www.sciencedirect.com/science/article/pii/S0889158314000422.

Wei, William WS. 2006. Time series analysis. *In: The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Wu, Bichen, Iandola, Forrest N., Jin, Peter H., & Keutzer, Kurt. 2016. SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving. *CoRR*, **abs/1612.01051**. http://arxiv.org/abs/1612.01051.

Yule, G Udny. 1926. Why do we sometimes get nonsense-correlations between Time-Series?–a study in sampling and the nature of time-series. *Journal of the royal statistical society*, **89**(1), 1–63.