International Journal of Computer Applications (0975 - 8887)
Volume 175 - No.27, October 2020

Machine Learning to Estimate the Floating Population in Florianópolis

Denilton Luiz Darold
Faculty of Business and Economics
Santa Catarina State University
Florianópolis, Brazil

Carlos Roberto Da Rolt
Faculty of Business and Economics
Santa Catarina State University
Florianópolis, Brazil

Andrea Sabbioni
Dpt. of Computer Science and Engineering
University of Bologna
Bologna, Italy

ABSTRACT

Touristic cities experience high fluctuation in their population, especially during the summer season. For many cities and countries, tourism plays a vital role in the economy, generating revenue and creating jobs. However, this so welcome economic boost comes along with an overload on public services, once the population increases dramatically in the high season. Therefore, an accurate method to predict the touristic demand is critical to provide the city administrators the necessary information for proper planning. Moreover, the private sector depends on demand forecasting to invest and maximize its profits. The most used methods currently rely on surveys and traditional indicators like the hotel's occupancy rates and arrivals at airports. Those methods, although assertive, miss relevant scenarios like the homestay and renting marketplace websites. This paper analyzes the use of machine learning techniques to predict the touristic load and, thus, the floating population in the Florianopolis' municipality year-round. It was analyzed several data sources, such as water consumption and international arrivals and its correlations to economic indicators like GDP, currency variation, temperature, and web search traffic. The results indicate that machine learning techniques applied to novel datasets pose great potential for achieving accurate forecasts.

Keywords

Floating population, seasonality, tourism measurement, machine learning

1. INTRODUCTION

The tourism sector is among the ones with the fastest global growth, showing increasing importance in many regions with economic and cultural impacts on the local communities. If the increase in Tourism flux can create new possibilities, it can also present new problems and challenges for the areas involved [12].

One of the main macro-effect caused by tourism is the steep growth of the population in specific areas for a relatively short period. This increase could result in undesirable side-effects, such as:

—Increase in air pollution due to the higher vehicles number and the concentration of traffic;

—Increase in localized crime;

—The overcrowding of community services like parking, public transport, museums, and rubbish bins;

—Increase in resource usages like water, gas, and electricity [1].

Sustainable planning involving local agencies and utilities aims to maximize tourism's positive aspects and minimize the adverse ones [7].

Since the fluctuation is temporary, the response actions must be taken in a fast and scalable fashion. Therefore, these solutions depend on an effective way to foresee future occurrences and their impact.

However, the measurement of the number of visitors is not a trivial task and, it varies greatly according to the region considered. The classical estimations methods are based upon surveys and data collected by government agencies that can present unreliable results due to scarcity or low quality of data.

The most commonly used indicators, like arrivals at airports and hotel's occupancy rate, disregard the occasional stay, i.e., the unregistered stay on private houses, like family houses or rented properties in web platforms. In a study conducted in 2007 in Florianópolis, the results obtained using the waste collection indicate a floating population three times higher than the outcome reached by the official bureau [3].

Given the pervasiveness of new technologies and how they are radically changing our behavior and lifestyle, the new concept of Smart Tourism is gaining more and more relevance.

The concept of Smart Tourism aims to exploit the massive amount of data produced and new possibilities enabled by the new ICT technology to create a more sustainable and accessible way of enjoying Tourism while valorizing the local cultural heritage and creativity. In particular, the concept of sustainability acquires the meaning, not only of a lower environmental impact but also the aim to create an offer that can positively support the local economy [9].

This study aims to create a predictive model able to estimate the populations' fluctuations in the particular case of the city of Florianópolis. To overcome the problems related to the accurate estimations of historical data on the population, it was conducted

a study over different datasets obtained by the ParticipACT project.

The ParticipACT project, started in the city of Bologna as a crowdsensing experiment, aims to study urban problems applying data-driven solutions[5]. ParticipACT Brazil is an extension of the original crowdsensing platform, including several research lines about the most pressing local issues, including mobility and the floating population. The project was also supported by local utilities like waste collection, electricity, water companies, and the city hall who cooperated by making many datasets available.

Thanks to this preliminary study and the data gathered interbred with an in-depth knowledge of the territory, it was possible to estimate which parts of the datasets are relevant for constructing the forecasting model.

The forecasting model was built with a Multiple Linear Regression implementation, considering the water consumption indicator as the dependent variable and the explanatory ones, the web search traffic(Google Trends), and predicted average maximum temperature. The predicted water consumption serves as a proxy to the proposed population estimation method along with a seasonality index. The period of analysis comprehends the years from 2013 to 2018.

The structure of the paper is as follows: In section 2, the background and local context are described. In section 3, related work is reviewed, while in section 4, the methodology is approached. Section 5 holds the details about the data used in the model proposed in section 6. The proposal is, thus, presented in section 6, while the results obtained and concluding remarks are in section 7.

2. BACKGROUND

The scope of this study is the Florianópolis municipality, the capital of Santa Catarina state, located in the south Brazil region. The estimated population, according to the Brazilian Geography and Statistics Institution (IBGE), was 469.690 inhabitants in 2015 [8]. Most of its territory (97.2%) is on the Santa Catarina Island, and the remaining area is on the mainland and surrounding small islands.

Three bridges link the island to the continent, but Hercílio Luz Bridge is presently closed for traffic. It is known that more than one hundred thousand people use the bridges every day mainly to work, but also to have access to public services, once the majority of public offices are located on the island [11].

The economy is mostly based on tourism, the public sector, and IT. The growth of the information technology sector in the city has encouraged the creation of technological parks and business condominiums to attract companies. Events related to information technology and sports are also important economic activities throughout the year.

In the next section are presented previous studies, mainly those with the same geographic scope.

3. RELATED WORK

In the field of Demographics and Tourism Measurement, several studies have already been done with the same aim as presented in this work, i.e., estimating the floating population through symptomatic variables. This section contains the most relevant ones on the development of the method created in this paper.

Reference [4] proposes water and energy consumption as symptomatic indicators to calculate the floating population in Sao Paulo State municipalities. A previous study is cited in the same study considering the hotel's accommodation capacity combined with vacant homes in a determined area. The vacant homes are multiplied by the residents' average by a house and then added to the hotel bedrooms number.

Reference [3] estimates the floating population through the waste production data and the number of connections to the energy grid.

Reference [6] made a time-series analysis of waste production, considering the seasonality. The author also mentioned the possibility of Water consumption, with no implementation due to the lack of data.

Reference [2] uses water consumption as an indicator. The difference to [4] lies in the assumption that the overall consumption metric is proportioned to the number of people, regardless of the temperature. However, this variable influences consumption. For instance, the energy is deeply impacted by the intensive use of air-conditioning systems during the summertime.

Reference [11] presents a comprehensive study using waste generation as a proxy variable. In particular, this study contributes to understanding the technological challenges of data preparation and the overall circumstances of this kind of research.

This paper proposes a new method, based on the previous studies, but with new data sources, novel technology, and the expertise of Water Company staff to determine the seasonal impact through the consumption patterns analysis. Regarding the data sources, it was added a near-real-time data source, the web search traffic. Lastly, the machine learning implementation presents a technological novelty, once the previous works were constrained to the statistics domain.

4. ANALYSIS METHODOLOGY

The methodology is divided into two parts: 1) The forecast of a dependent variable through multiple linear regression, and 2) the actual floating population estimation.
In phase 1, a multiple linear regression takes place, summarizing the statistics in order to identify and evaluate the indicators, its correlation, and the linearity to ensure that the Linear Regression is, indeed, an adequate approach. For this purpose, the following steps were taken:

—Analysis of descriptive statistical indicators;

—Removal of the outliers;

—Analysis of correlation;

—Creation of the initial prediction model;

—Testing the model, review, and simulate scenarios.

Once phase 1 is concluded, and the predictive model is ready, it is possible to foresee the dependent variable, in our case, the water consumption. As previously mentioned, this indicator serves as a proxy to calculate the population projections in phase 2.

The methodology applied in this second phase is a combination of the methods mentioned in section II, especially the one approached in the [6]. The overall logic is to find out a seasonality coefficient and multiply it by the resident's number. In order to accomplish that, were necessary the following steps:

—Creation of a new dataset with twelve dummy variables, representing the months of the year;

—Calculation of seasonality coefficient;

—Choice of a touristic neighborhood;

—Mitigating the influence of external factors, mainly the temperature, once the individual water consumption increases in the summertime;

—Estimate the population by applying the coefficient to the official resident number.

5. DATA SOURCES

The datasets used in this paper were selected based on the previous related work and tourism-related researches in general.

The data were obtained mainly from the research project ParticipAct, which collects data from public sources to conduct data analysis regarding urban problems.

As the Florianópolis districts present a heterogeneous touristic demand, the analysis was restricted to a single neighborhood with high touristic demand, namely Canasvieiras. The data used refers to the 72 months from 2013 to 2018. The datasets considered in this paper were:

—Water Consumption: Dataset kindly provided by the state-owned water company CASAN. The data indicates the monthly consumption measured, by street, in cubic meters;

—Web Search Traffic: The data was obtained from the web application Google Trends and represented the overall interest of a topic, in this case, the Florianópolis City. Still, it was collected the indicator, both with origin limited to Argentina, the neighboring country, as global;

—Energy Consumption: Dataset obtained on the Data Catalog of ParticipAct project through a technical partnership with the Santa Catarina Power Plants company, CELESC. The data indicates the monthly consumption measured, by street, in kWh;

—Waste production: Dataset obtained on the Data Catalog of ParticipAct project through a technical partnership with the Waste collection company, COMCAP. It Indicates the amount of waste collected, by route, in tons. The indicators by the route and not street pose a challenge once the routes can stretch to more than one neighborhood;

—Temperature: The average maximum temperature forecast in Celsius degrees. The data was collected from the National Institute of Meteorology, INMET;

—GDP: The Gross Domestic Product variation in %, collected from Brazilian Geography and Statistics Institution (IBGE);

—Currency variation (Both to US dollars and Argentinian Pesos, ARS) in %, collected from website Investing.com.

The water consumption data was chosen as the dependent variable after been mentioned and recommended in the previous related work. In [6], mentioned but not utilized due to the lack of data. In [10], the water was considered the most satisfactory indicator, mainly due to its linearity. The comparison between different municipalities shows a direct proportion between consumption and population.

Another relevant dataset included in this study is the web search traffic, aforementioned, Google Trends. This dataset represents the volume of queries made on Google in a period. It was considered the global index and a filtered one, limiting the origin of the search to Argentina, our neighboring country, and the most significant visitor. The proportion of Argentinians is depicted in Fig. 1, through the tourist's country of origin statistics, collected from the Brazilian Tourism Ministry, MTur.

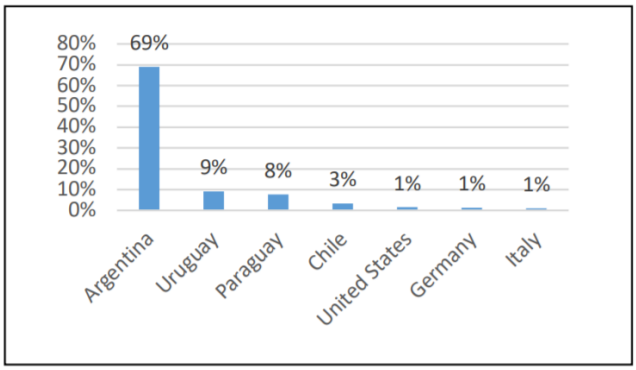

Fig. 1. Percentage of Tourists Residence Country 2014-2018

6. PROPOSAL

In this section, the predictive model is presented. For this purpose, datasets from different sources were consumed, as previously mentioned. As described in section 3, the method proposed is to use symptomatic indicators to train a model to predict future values based on a Multiple Linear Regression. The variable chosen as the dependent one (y) was water consumption. The reason is the water consumption presents a better statistical significance [10]. Therefore, the water consumption estimation it will serve as a proxy to calculate the floating population, according to a seasonality index, as approached further in this section.

This section contains Descriptive statistics, Linear Regression, Implementation, and finally, Floating population estimation.

6.1 Descriptive Statistics

In the descriptive statistics, it is possible to quickly assess the data's consistency through the standard deviation, mean, minimum, and maximum values. Considering the data analyzed was aggregate by month, no data transformations were needed.

Afterward, a Pearson matrix correlation was generated. The correlation index varies from -1 to 1, been negatively correlated and positively correlated, respectively. Table I indicates a positive correlation between Water, Forecasted Temperature, and Google

Trends Argentina. The Google Trends Argentina presented a moderate positive correlation. The other indicators, GDP variation, Currency variation, presented a low or negative correlation and, therefore, were removed from the predictive model. The Energy and Waste collection were also removed due to multicollinearity, i.e., they represent the same phenomenon.

Table 1. Pearson Correlation Matrix

	Water	Temperature [a]	Trends [b]
Water	1.0000	0.7933	0.8020
Temp	0.7933	1.0000	0.7560
Trends	0.8020	0.7560	1.0000

[a] Forecasted Average Maximum Temperature (C°)
[b] Google Trends from Argentina

An additional way to portrait the correlation and linearity of the dependent and explanatory variables is to render scatterplots, as shown in Fig. 2 with the average temperature and Google Trends related to water.

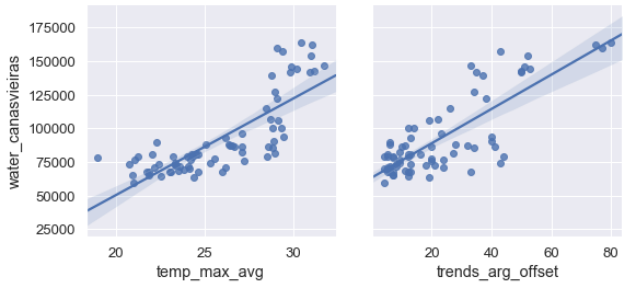

Fig. 2. Scatterplot water vs predicted temperature and Google Trends

6.2 Linear Regression

Linear regression models have been used to study several research problems in many different fields, including tourism. They are broadly used in the literature to examine tourism demand, and within the gamma of estimation methods, the classical Ordinary Least Square(OLS) is regularly employed [11].

Conceptually, the Linear Regression refers to the study of the dependence of a variable on one or more variables, the explanatory variables, aiming to estimate or predict the mean values of the former in terms of known values of the latter.

The Linear Regression can be expressed by the following formula (1):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \quad (1)$$

Where: TODO
$\hat{Y}_i$: The dependent variable. In this study, the amount of water consumption.
$\hat{\beta}_0$: Intercept
$\hat{\beta}_1$: Coefficients of Regression.
Xi: Independent variables, Google Trends Argentina and Average Maximum Temperature, respectively

6.3 Implementation of the Predictive Model

This section presents a glimpse of the implementation of the solution. The predictive model was built using a Python module, the scikit-learn[1], well-known in the data science community. This module provides several machine learning methods, like classification, regression, and clustering. In this paper, the regression was utilized for the reason already pointed out.

Initially, it was defined the dependent variable and the explanatory ones. Then, it was set the split proportion to create a subset to train and others to test our prediction, as seen on Listing 1. The test dataset size was set to 0.33, approximately two-thirds, corresponding to two of six years of dataset timespan. The other four years, were, therefore, used to train the model.

```
#Importing library
from sklearn.model_selection import train_test_split

#creating a data Series of water consumption
y = data['water']

#creating a data frame with explanatory variables
X = data [['temp_max_avg', 'trends_arg_offset']]

#creating the train and test datasets
X_train, X_test, y_train, y_test = train_test_
split(X, y, test_size=0.33, random_state=2811)
```

Listing 1 - Spliting of Train and Test sets

Afterward, it is time to finally train the model, as shown in Listing 2.

```
#Importing libraries
from sklearn.linear_model import LinearRegression
from sklearn import metrics

#Creating the Model instance
model = LinearRegression()

#Training the model
model.fit(X_train, y_train)
```

Listing 2 - Model training

The coefficient of determination, R^2, obtained was 0.71 to the train dataset and 0.72 to test dataset, on a scale from 0 to 1 and indicates how well our model predicts the water generated in Florianópolis. This coefficient is commonly used to represent how the regression line fits the data. An R^2 of 0.50, for instance, indicates that 50 percent of the variation in the result has been explained by the covariates included in the model. It's defined as $1 - u/v$, where u is the residual sum of squares $((y\_true - y\_predicted) * *2).sum()$ and v is the total sum of squares $((y\_true - y\_true.mean()) * *2).sum()$ [90].
Another metric commonly considered is the mean square error, which corresponds to the expected value of the squared (quadratic) error or loss. In this model, the outcome was 7.000000e-01.

[1] https://scikit-learn.org/stable

To evaluate the variables used in the model, individually, it is used the intercept, and the regression coefficients, according to the linear regression formula aforementioned. The intercept represents the mean effect on Y, excluding all variables from the model. In our model, that means that the effect on water consumption when Temperature and Google Trends Argentina were zero. The outcome was -27137.27 m^3.

The regression coefficients are known as partial regression coefficients or angular coefficients. Given the explanatory variables of our model, their meaning would be the measured variation in the mean value of Y(Water) by variation unit in X2(Temperature), keeping the values of X3(Trends) constant. In an analog way, it is possible to interpret the other regression coefficient, X3.

At this point, the model is considered trained, verified, and ready to predict the water consumption given a Temperature and Google Trends Argentina. Through the outcome values, it is possible to estimate the floating population.

6.4 FLOATING POPULATION

In this section, is conducted a monthly estimation based on the outcome of the proxy variable, namely, the water consumption projection, the seasonality index, and official census population data. The period aimed was the year 2017.

The first step is to create a new dataset on a monthly basis, with all indicators. Afterward, twelve dummy variables were added, one to each month. The criteria of values were as follows:

—Dummy variable equals 1 if the record is from the same month represented by variable;

—Dummy variable equals 0 if the record is not from the same month represented by variable;

The next step is to calculate the monthly coefficient. For this purpose, it was used the statistic tool Gretl[2]. The result is showed in Table II, with the Ordinary Least Squares (OLS) indicators. The Gretl removed the month of December automatically due to collinearity.

Table 2. Monthly OLS

	coefficient	error	t-test	p-value
const	86384	3.120	27,690	7,1E-22
Jan	61225,5	4.412	13,878	2,2E-06
Fev	66017,3	4.412	14,964	6,5E-08
Mar	25944	4.412	58,805	1,9E+07
Apr	5237	4.412	11,870	2,4E-01
May	−6595,83	4.412	−14,950	1,4E−01
Jun	−13220, 5	4.412	−29, 966	4,0E-03
Jul	−16551	4.412	−37, 515	4,0E-04
Aug	−14796, 7	4.412	−33, 538	1,4E-03
Sep	−12974, 2	4.412	−29, 408	4,6E-03
Out	−13918, 3	4.412	−31, 548	2,5E-03
Nov	−5827, 83	4.412	−13, 210	1,9E-01

Through the monthly coefficients, it is possible to determine a seasonality index. The method proposed is to divide the monthly coefficient by the constant-coefficient value in the first row. The lowest coefficient represents the month with the lowest consumption, and then, presumably, the month in which only residents are actually living in the neighborhood, in this case, August, with a population of 20.017 inhabitants.

Another factor taken into consideration was the natural increase in water consumption during the summer season. According to Water Company, CASAN, this seasonal increase represents approximately 6% of total consumption. This percentage was obtained comparing the consumption patterns between touristic and non-touristic cities throughout the year. In this analysis, the summer months in touristic cities presented an increase equivalent to two standard deviations, which by the company statistics represents 6 percent on consumption.

Considering the seasonality index and the discount of natural increase in the summer months, namely January, February, and Mars, the outcome is shown in Table III.

Table 3. Floating Population Estimation

Month	Seasonality index	Population	Floating population
Jan	1,702715963	40.727	19.468
Feb	1,762608246	42.160	20.814
Mar	1,32406279	31.670	10.954
Apr	1,058633587	25.321	5.304
May	0,916320126	21.917	1.900
Jun	0,835355228	19.981	-36
Jul	0,836866712	20.017	-
Aug	0,852436567	20.389	372
Sep	0,883468727	21.132	1.115
Out	0,865747837	20.708	691
Nov	0,972941112	23.272	3.255

Figure 3 depicts the monthly variation of the population in the Florianopolis' neighborhood of Canasvieiras, indicating the February as the month with the higher population and August with the lowest population

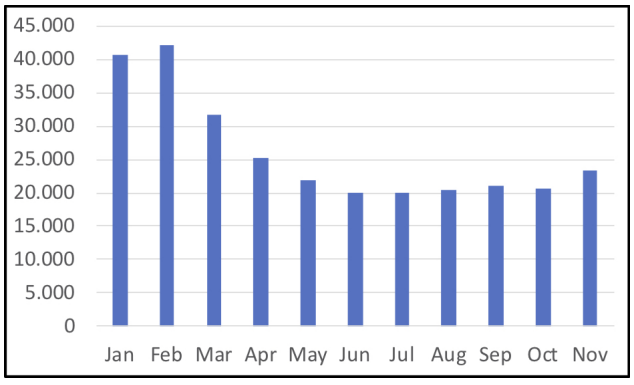

Fig. 3. Floating Population variation

The highest number in February could be explained, besides the seasonality factor, due to Brazil's prolonged holidays during the Carnival period, which takes place usually in that month.

[2]Gretl: Gnu Regression, Econometrics and Time-series library

7. CONCLUSION AND FUTURE WORKS

This paper lays the foundations for the use of machine learning techniques for the floating population estimations. The model hereby proposed, combining near real-time Google web search information along with public databases, provide stakeholders a new set of tools to model complex data analysis such as floating population.

During the data analysis and model development, was possible to validate the following hypothesis: a) The multiple linear regression method is adequate to estimate water consumption, due to the linear relationship of the dependent and explanatory variables; b) The economic indicators, like GDP and Currency variation, has no positive correlation and, therefore, do not explain the floating population; c) It is viable to create a machine learning model to predict water consumption and, finally, d) It is possible to estimate the floating population through a proxy variable.

The findings of the present work can help public managers, researchers, and even companies in the tourism sector build their estimations and, thus, analyze the dynamics of the population. The knowledge can be useful in different domains and to different stakeholders. It can be applied to the public sector in resource planning, and therefore, better services provision. To the private entities, to support them to invest wisely to deliver better services to visitors. Tourism is one of the most significant revenue sources, and the cities should be prepared to provide visitors with excellent public and private services.

As highlighted in this paper and other studies in different regions, one of the main problems in pursuing an accurate predictive model in tourism demand is the lack of geolocated data to a detailed temporal analysis.

The integration with a crowdsensing infrastructure based on IoT devices can enable a more precise and fine-grained data collection that improves the forecasting model's accuracy.

Moreover, the exploitation of new data sources, like social media, and other methods like websites crawlers and scrapers, can undoubtedly be considered to support the forecasting model. This development can also enable real-time and district-based forecasting, enabling even more advanced and dynamic solutions to the authorities.

Finally, this could highlight that a model for forecasting tourism afflux varies according to the local context. Also, the circumstances and interests change over time, affecting the tourism demand. For that reason, the development of advanced machine learning models capable of understanding and self-adapt to new scenarios is a promising research field.

8. REFERENCES

[1] M. Novelli e J. M. Cheer C. Milano. Overtourism and Tourismphobia: A Journey Through Four Decades of Tourism Development, Planning and Local Concerns. 2019.

[2] Andrés Camacho Murillo. Methods and measurement techniques in tourism. *Turismo y Sociedad*, 24(November):211–216, 2018.

[3] Paulo Campanario. Florianópolis: dinâmica demográfica e projeção da população por sexo, grupos etários, distritos e bairros (1950-2050). *Ipuf*, 2007.

[4] Rute Eduviges Godinho. Nova metodologia de projeção da população flutuante. 1:1–13, 2000.

[5] Eliza Gomes, M. A.R. Dantas, Douglas D.J. De Macedo, Carlos De Rolt, Marcelo Luiz Brocardo, and Luca Foschini. Towards an infrastructure to support big data for a smart city project. *Proceedings - 25th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2016*, pages 107–112, 2016.

[6] Antonio Guarda. Gestão Urbana : Projeção da População Flutuante Gestão Urbana : Projeção da População Flutuante. (January 2012), 2015.

[7] M. de la Calle-Vaquero e C. Yubero M. García-Hernández. Cultural heritage and urban tourism: Historic city centres under pressure, Planning and Local Concerns. 9(8):4–8, 2017.

[8] Brazilian Institute of Geography and Statistics. Cities and States. `https://www.ibge.gov.br/cidades-e-estados/sc/florianopolis.html`, 2019. [Online; accessed 28-June-2019].

[9] European Capital of Smart Tourism Secretariat. European capital of smart tourism, Oct 2018.

[10] Pedro Tonon Zuanazzi and Mariana Bartels. Estimativas para a população flutuante do Litoral Norte do RS. page 29, 2016.

[11] Sergio Dalla Valle. *Metodologie di integrazione di crowdsensing e big data dal territorio : l ' esperienza Participact Brazil per smart city*. PhD thesis, UNIBO, 2018.

[12] WTTC. Travel and tourism economic impact 2018 world, Oct 2018.