UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

FELIPE GIOACHINO OPERTI

COMPUTATIONAL ANALYSIS FOR SOCIO-ECONOMIC SCIENCES

FORTALEZA

2018

FELIPE GIOACHINO OPERTI

COMPUTATIONAL ANALYSIS FOR SOCIO-ECONOMIC SCIENCES

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Física da Universidade Federal do Ceará, como requisito parcial para a obtenção do Título de Doutor em Física. Área de Concentração: Física da Matéria Condensada.

FORTALEZA
2018

FELIPE GIOACHINO OPERTI

COMPUTATIONAL ANALYSIS FOR SOCIO-ECONOMIC SCIENCES

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Física da Universidade Federal do Ceará, como requisito parcial para a obtenção do Título de Doutor em Física. Área de Concentração: Física da Matéria Condensada.

Aprovada em 17/12/2018.

BANCA EXAMINADORA

Dr. José Soares de Andrade Jr. (Supervisor)
Universidade Federal do Ceará (UFC)

Dr. Humberto de Andrade Carmona (Internal)
Universidade Federal do Ceará (UFC)

Dr. André Auto Moreira (Internal)
Universidade Federal do Ceará (UFC)

Dr. Joao Jose Vasco Peixoto Furtado (External)
Universidade de Fortaleza (Unifor)

Dr. Haroldo Valentin Ribeiro (External)
Universidade Estadual de Maringá (UEM)

To my parents and grandparents, for those who are still here and for those who left.

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O1c Operti, Felipe Gioachino.
COMPUTATIONAL ANALYSIS FOR SOCIO-ECONOMIC SCIENCES / Felipe Gioachino Operti. – 2018.
111 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Física , Fortaleza, 2018.
Orientação: Prof. Dr. José Soares de Andrade Jr..

1. Complex Systems. 2. Data mining. 3. Economic Complexity. 4. Racial Residential Segregation. 5. Allometry. I. Título.

CDD 530

ACKNOWLEDGMENT

First and foremost I would like to thank my advisor Professor José Soares de Andrade Jr. for the help and the support. I also thank my collaborators in Rome, especially the PhD Andrea Gabrielli and in New York City, the Professor Hernan Makse. Furthermore, I thank the CAPES, CNPq, and FUNCAP for the financial support.

I really thank my mother and my father for their support. They supported me in all my choices with respect. I also thank my family both Italian and Brazilian for the help and the love.

A special thank my friends both Italians and Brazilians such as Samuel, Rilder, Joao, Nena, Aurelio, Tatiana, Saulo, Eduardo, Israel, Cesar, Cezinha, Hygor, Erneson, Marciel, Thiago, Heitor, Wagner, Nicolò, Simone, Alessandro P, Alessandro R, Davide, Marco, Maurizio, and many others.

ABSTRACT

In this thesis we show three projects in the field of the physics of complex systems. All of them are based in the statistical analysis of real data, subfield nowadays commonly known as Data Science. Indeed, the last years have been characterized by a rapid growth of the amount of data. These data range from economy to biology, from finance to astrophysics, and many others. Generally the extraction of the information from them is a complex problem and it requires advanced statistical and mathematical frameworks. Here, we show three projects which are based in the statistical analysis of real data. In the first, entitled *The light pollution as a surrogate for urban population of the US cities*, we approached the problem of the light pollution analyzing the scaling of the population of the US cities with the night-time light. In the second project, entitled *Dynamics in the Fitness-Income plane: Brazilian States VS World Countries* we provide a variant of the Fitness algorithm (a novel method to compare the economic development of world countries) to measure the development of the Brazilian states. In the third and last project, entitled *Dynamics of racial segregation and gentrification in New York City* we focused in the analysis of the racial residential segregation. In that project we introduce a new index of segregation and we compare the racial residential segregation with several other factors such as the per capita income, the properties values, and flux of people, finding connections with the gentrification of some neighbors of New York City. We conclude the thesis providing the main results of each project and emphasizing the importance of the interaction among scientists of different areas in the study of socio-economic sciences.

Keywords:Complex Systems, Data mining, Night-time light, Light Pollution, Allometry, Fitness, Complexity, Economic Complexity, Racial Residential Segregation, Gentrification, Inequalities.

LIST OF FIGURES

SUMÁRIO

1 INTRODUCTION

The amount of open data is growing exponentially in the last twenty years. Indeed, according with Barnard Marr of Forbes [1] during the last year alone, it was generated the 90 percent of the data in the world. These data regards different areas, from finance and economics, to astrophysics, social networks, biology and many others fields. The analysis and the extraction of information from them could be a very complex problem and it begun source of interest for scientists in different areas. In this context, the physicist of complex systems has found a fertile soil.

In this thesis we show three different research projects, two of them are already published in international journals, while one is currently submitted. The field of study differs from project to project, indeed they vary from the economy to the study of the racial residential segregation and to the allometric analysis of the night-time light. However, these different fields of study are brought together by the methodologies used to analyze the large amount of real data of which they are supplied. In fact, the aim of the data scientist is to analyze the real data using mathematical frameworks such as, for example, stochastic processes, statistics, or, latest methods such as machine learning, deep learning, and neural networks.

The aim of this thesis is to show three research projects in the area of the complex systems. The thesis is structured as follows: in the Chapter two entitled *Basic notions of Data Science*, first we introduce the process of *data mining*. This part is the ground for all the projects, and show the types of data used in this thesis. Second, we provide a brief review about the *Science of Cities*, emphasizing the meaning of *allometric scaling* and the *City Clustering Algorithm* (an algorithm to define the limit of the urban centers). Third we introduce a new branch of the econophysics called *Economic Complexity*. The aim of this chapter is to provide a first approach to the statistical analysis of real data focused in the projects discussed in the next chapters. However, each successive chapter is provided of a more detailed introduction for the specific field covered.

In the Chapter three, entitled *The light pollution as a surrogate for urban population of the US cities*, we show that the definition of the city boundaries can have a dramatic influence on the scaling behavior of the night-time light (NTL) as a function of population (POP) in the US. Precisely, our results show that the arbitrary geopolitical definition based on the Metropolitan/Consolidated Metropolitan Statistical Areas (MSA/CMSA) leads to a sublinear power-law growth of NTL with POP. On the other hand, when cities are defined according to a more natural agglomeration criteria, na-

mely, the City Clustering Algorithm (CCA), an isometric relation emerges between NTL and population. This discrepancy is compatible with results from previous works showing that the scaling behaviors of various urban indicators with population can be substantially different for distinct definitions of city boundaries. Moreover, considering the CCA definition as more adequate than the MSA/CMSA one because the former does not violate the expected extensivity between land population and area of their generated clusters, we conclude that, without loss of generality, the CCA measures of light pollution and population could be interchangeably utilized in future studies.

In the Chapter four called *Dynamics in the Fitness-Income plane: Brazilian states vs World countries*, we introduce a novel algorithm, called Exogenous Fitness, to calculate the Fitness of subnational entities and we apply it to the states of Brazil. In the last decade, several indices were introduced to measure the competitiveness of countries by looking at the complexity of their export basket. Tacchella et al (2012) developed a non-monetary metric called Fitness. In this project, after an overview about Brazil as a whole and the comparison with the other BRIC countries, we introduce a new methodology based on the Fitness algorithm, called Exogenous Fitness. Combining the results with the Gross Domestic Product per capita (GDP_p), we look at the dynamics of the Brazilian states in the Fitness-Income plane. Two regimes are distinguishable: one with high predictability and the other with low predictability, showing a deep analogy with the heterogeneous dynamics of the World countries. Furthermore, we compare the ranking of the Brazilian states according to the Exogenous Fitness with the ranking obtained through two other techniques, namely Endogenous Fitness and Economic Complexity Index.

In the Chapter five entitled *Dynamics of racial segregation and gentrification in New York City*, we developed a new method in order to measure and to define the topography of racial residential segregation. Racial residential segregation is interconnected with several other phenomena such as income inequalities, property values inequalities, and racial disparities in health and in education. Furthermore, recent literature suggests the phenomena of gentrification as a cause of perpetuation or increase of racial residential segregation in some American cities. In this project, we analyze the dynamics of racial residential segregation for white, black, Asian, and Hispanic citizens in New York City in the years of 1990, 2000, and 2010. It was possible to observe that segregation between white and Hispanic citizens, and discrimination between white and Asian ones has grown, while segregation between white and black is quite stable. Furthermore, we analyzed the per capita income and the Gini coefficient in each segregated zone, showing that the highest inequalities occur in the zones where there is overlap of high-density zones of pair of races. Focusing on census tracts that have changed density of population during these

twenty years, and, particularly, by analyzing white and black people's segregation, our analysis reveals that a positive flux of white (black) people is associated to a substantial increase (decrease) of the property values, as compared with the city mean. Furthermore, by clustering the region of high density of black citizens, we measured the variation of area and displacement of the four biggest clusters in the period from 1990 to 2010. The large displacements ($\approx 1.6\ km$) observed for two of these clusters, namely, one in the neighborhood of Harlem and the other inside the borough of Brooklyn, led to the emergence of typically gentrified regions.

Finally, in the Chapter six, we provide the *General conclusions*. Furthermore, in the appendix (*Appendix A*) we attached the copies of the international publications of the first two projects.

2 BASIC NOTIONS OF DATA SCIENCE

To make clear the results showed in this thesis some theoretical frameworks are necessary and they are partly shown in this Chapter and partly shown in the section *Introduction* of each chapter. As it was introduced in the previous Chapter the common ground among these projects is the methodology to analyze the data. In this context, the first section is focused on the *data mining* process.

2.1 Data mining

The process starts with the acquisition of the data. There are many types of data and formats, each one with a specific function. In this thesis we used two types of data: text data (where the information of interest are enclosed), and shapefile data (where the geometric and geospatial information are enclosed). However, text data could be provided in different formats such as csv, json, txt, ascii, xml, graphml, etc. More information about the data acquisition is attached in each chapter.

The amount of data are often very big and, sometimes, a powerful hardware are needed to processing of them. Moreover, the software tools are extremely important to lead with the major part of the data. The most part of the results showed in thesis are obtained using R and Python as programming language combined with the libraries Pandas, Numpy, Sci-Kit learn, Matplotlib, and Seaborn. Furthermore, to the visualization of georeferenced data we used the software QGis.

Data mining is the process of learning from data using mathematical frameworks such as statistics, artificial intelligence, machine learning, deep learning, and neural networks [2]. In recent years, the use of machine learning algorithms has spread among the scientists. Actually some techniques have become famous as machine learning algorithms were already known and commonly used by statistical and mathematical physicists [2]. Among them, probably the most widespread is the linear regression.

Machine learning is divided in two big categories: *supervised learning* and *unsupervised learning* [2]. In supervised learning are enclosed all the algorithms that use input data as well output data. To be more clear, the aim of the supervised algorithms is to find the rules and laws that exist (if they exists) from an input data to an output data. While the aim the unsupervised algorithms is to find possible relations inside a group of data [2]. In this thesis we use both supervised and unsupervised algorithms.

In the next sections, we introduce two fields main topics of the next chapters,

respectively called *The Science of Cities* and the *Economic Complexity*.

2.2 The Science of Cities

The book *The new science of cities* written in 2013 by Michael Batty [4] is the first book focused in the scientific study of the growing of the cities. In that book, M Batty [4] lists the foundations of this new science. The aim of the science of cities is the study and the analysis of the correlation between the growth of the urban centers (population or area) and other variables such as, for example number of crimes, income, jobs, number of suicides, gas emissions, light pollution, etc.

The study of the scaling of the cities with other socio-economic indicators is many times compared with the biological allometric study [5]. In the next subsection we deepen the history and the meaning of the allometry study of urban centers. However, not less important, is the definition of the limit of cities. In fact, several studies show the scaling of the cities with other variables deeply depends by the definition of the limit of the urban centers [6,7]. In this context, in this chapter we explain an algorithm developed with this goal, the *City Clustering Algorithm* [5–12].

2.2.1 Allometry

Allometry is a term introduced in biology between the end of the nineteenth century and the begging of the twenty century [13]. The term is first used to define the study of the relationship between the body size and shape/anatomy/physiology/behavior of the living organisms. This relationship is often expressed as a power law:

$$y = kx^a, \tag{2.1}$$

where x is the body size and y is another feature such as shape/anatomy/physiology/behavior. While a is called allometric exponent. The information of the scaling between the two quantities is totally incorporated in the allometric exponent a. Furthermore, the same expression can be visualized in logarithmic form:

$$log(y) = alog(x) + log(k). \tag{2.2}$$

Therefore re-defining $y' = log(y)$, $x' = log(x)$, and $k' = log(k)$, the power law showed in eq. 2.1 can be expressed as linear expression as:

$$y' = ax' + k'. \tag{2.3}$$

In this case the allometric exponent is the linear coefficient of the relation showed in eq.

2.3. While when the allometric exponent $a \approx 1$ the relation is called isometric. In fact in this case the eq. 2.1 is not a power law, but it is linear.

Bettecourt and his collaborators in 2007 [5] showed deep analogies between living organisms and cities. In this context they found the cities in the US exhibit three different types of allometric laws for urban indicators with population size [5]: (i) *Superlinear* . The superlinear urban indicators increase proportionally more than the population of the cities. Such behavior is intrinsically associated with the *social currency* of a city, indicating that larger cities are associated with optimal levels of human productivity and quality of life. Doubling the city size leads to a larger-than-double increment in productivity and life standards [5, 36, 37]. For example, wages, income, growth domestic product (GDP), bank deposits, as well as rates of invention measured by the number of patents and employment in creative sectors show a superlinear behavior [5]. (ii) *Linear* or *isometric relation*. The increasing of the linear urban indicators is proportional to the increasing of the population reflecting the common individual human needs, like the number of jobs, houses, and water consumption [5]. (iii) *Sublinear*. The sublinear urban indicators increase proportionally less than the population of the cities. This case is a manifestation of the *economy of scale*. The sublinearity is found in the number of gasoline stations, length of electrical cables, and road surfaces (material and infrastructure) cases [5].

Moreover, an allometric study published by Oliveira *et al* in 2014 showed as the definition of the limits of the cities affects the analysis [7]. In their papers they studied the scaling of the CO_2 emission in function of the population of the US cities. Although they used two methods to define the limits of the cities. In the next section we deepen one of these methods called City Clustering Algorithm (CCA). The same algorithm will be used in the next chapters of this thesis.

2.2.2 City Clustering Algorithm (CCA)

The CCA is an algorithm introduced to define boundaries of metropolitan areas [5–12]. Its result depends on two parameters, namely, a population density threshold, D^* and a cutoff length, ℓ [11]. For the i-th grid cell, the population density D_i is georeferenced in its geometric center (shown as small black circles in Fig. 1). If $D_i > D^*$, the i-th grid cell is populated. In Fig. 1 the populated cells are shown in grey and red. Next, the algorithm selects a populated cell (red cell in Fig. 1a) and aggregates in the same cluster all nearest populated cells which are within a distance ℓ from each (red cells in Figs. 1b, 1c and 1d). The Fig. 1 shows the four steps to determine the red cluster.

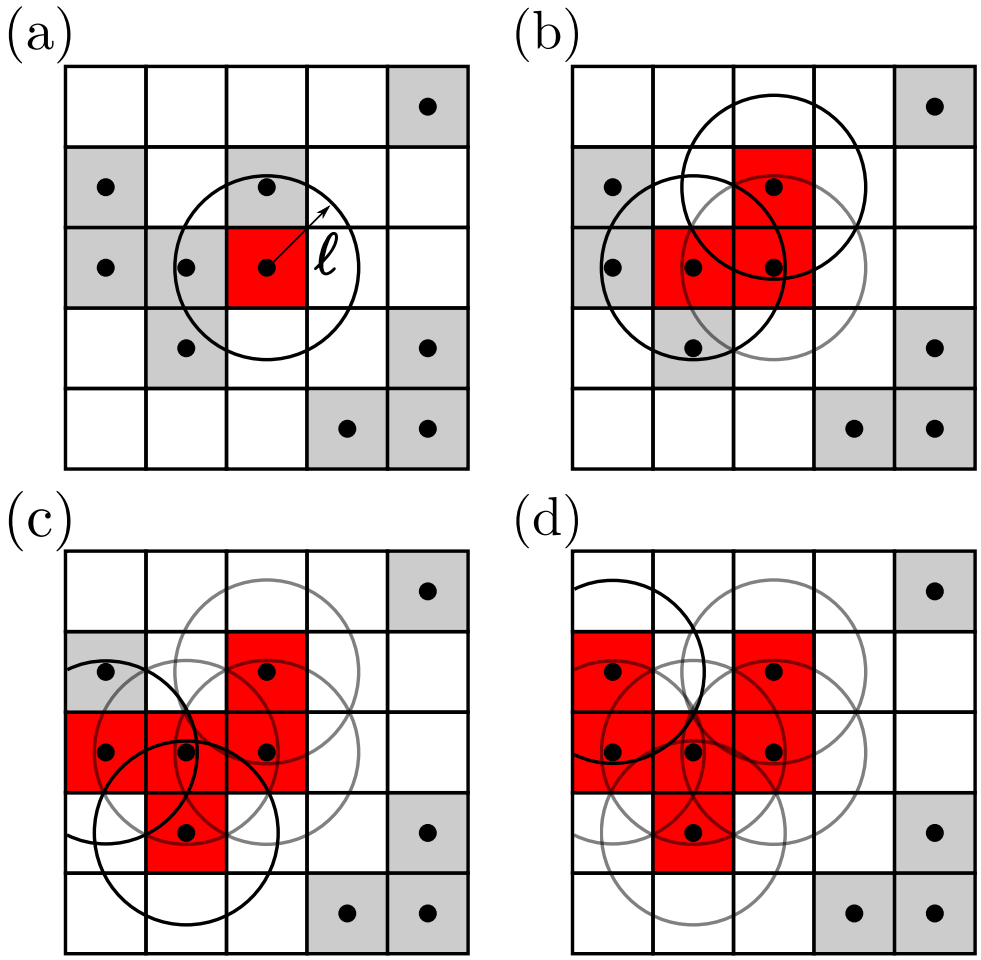

Figura 1: **The CCA steps (on colors).** The grey and the red cells are populated ($D_i > D^*$). The small black circles are the geometric centers of each populated cell. The red cells belong the same analyzed cluster. (a) First step: the algorithm select a populated cell and draw a circle of radius ℓ. (b) Second step: the cells with the geometric centers inside the circles of radius ℓ become a part of the red cluster and from their geometric center are drawn others two circles of radius ℓ. The circle of the first step is showed in opaque black. (c) Third step: two more cells became part of the red cluster and two more circles are drawn. (d) Fourth step: the last cell became part of the red cluster. The entire cluster is determined and the algorithm will start to analyze another cluster.

2.3 Economic Complexity

There is a tendency of the economists to focus on monetary indexes to analyze the world economy. Among them, the most used is the *Gross Domestic Product* (GDP). However, GDP alone, as shown by different studies [14–17], does not provide all the information about the perspective of growth and development of world countries, because it is strongly sensible by the inflation and the national currencies. This fact makes difficult to compare the economies of different countries and their evolution during the years. In this context, economists and more in general scientists, developed non-monetary indexes

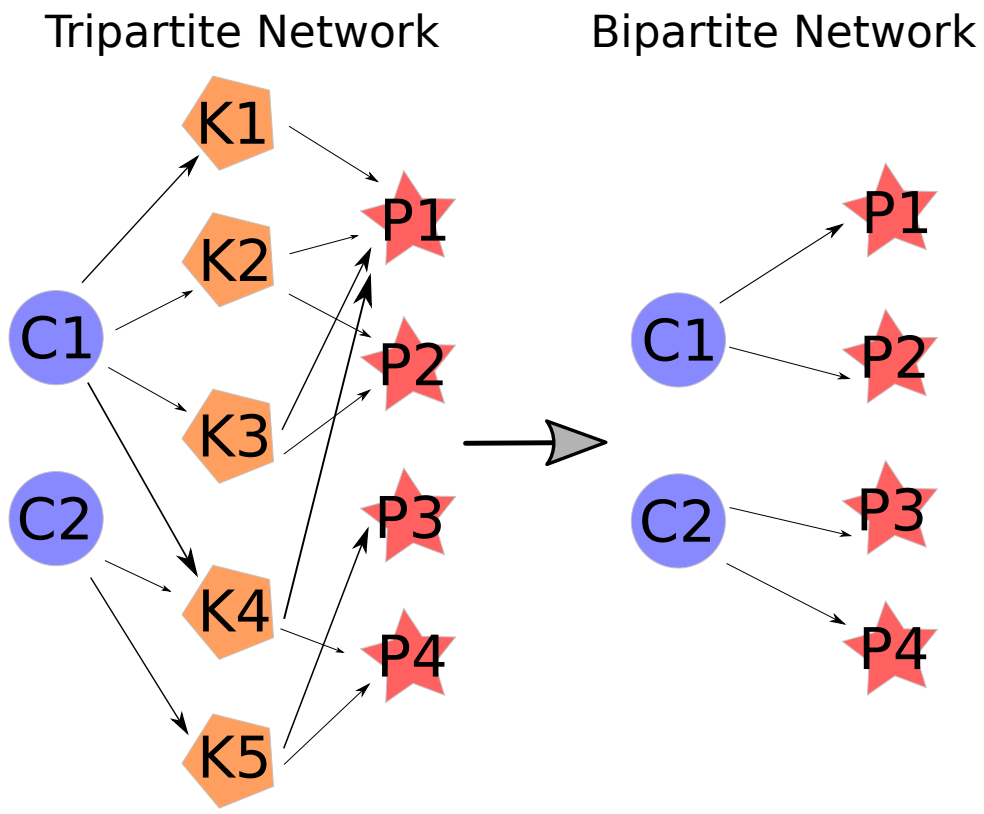

Figura 2: **From tripartite to bipartite.** The figure shows the tripartite network where countries (C1 and C2) are connected with the capabilities (K1, K2, K3, K4, and K5) that are connected with the products (P1, P2, P3, and P4). The tripartite network can be reduced in a bipartite network countries/products.

with the aim to quantify and compare the development of the world countries overcoming the faults of the monetary indexes [14–18].

With this aim, Ricardo Hausmann and Cesar Hidalgo [14] studied the economy as a tripartite network as showed in the Fig 2. The figure shows the world economy as a tripartite network where the countries are first connected to several "capabilities". The capabilities are the ensemble of features such as, for example, infrastructures or education, that a country needs to produce and to export a product. These features are hardly quantifiable, however in the tripartite network the capabilities are connected with the products. Indeed capabilities allow the production of each product and moreover their exportation. Nonetheless, capabilities are not quantifiable, therefore the tripartite network is reduced in a bipartite network countries/products where only the exported products are analyzed [14].

The bipartite network showed in Fig 2, inspired different authors in the formulation of new non-monetary indexes able to describe the development of world countries. The first, formulated in 2009 by R Hausmann and C Hidalgo [14] is called Economic Complexity Index (ECI). After that, the same concept has been improved and enriched

by Cristelli and his collaborators in 2012 and called Economic Fitness [15]. In this thesis, we deepen the Fitness and we extend it to subnational entities [18].

3 THE LIGHT POLLUTION AS A SURROGATE FOR URBAN POPULATION OF THE US CITIES

3.1 Introduction

More than 80% of the world and more than 90% of the US and European populations live under light-polluted skies (exposition to light at night) [26]. Since the first electric-powered illumination in the second half of the 19th century, the world has become covered by artificial electric light, changing drastically the night view of the Earth from space. The spreading of artificial electric light plays an important role on the duration of the *productive day*, not only for working but also for recreational activities. If in one hand the benefits of artificial light are quite evident, on the other hand, scientific researches suggest that the exposition to light at night could have adverse effects on both human and wildlife health [27–34]. For example, in humans, the pineal and blood melatonin rhythms are quickly disturbed by light pollution. Such studies argue that the night light exposure have two major physiological effects: they disrupt the circadian rhythms and suppress the production of melatonin [33]. This repeated suppression may have large consequences for the mammals health. For instance, it was shown that the suppression of the melatonin at night accelerates the metabolic activity and growth of rat hepatoma [30] and human breast cancer [28]. Moreover, the disruption of circadian rhythms made by the exposure of light at night might plays a crucial role in the etiology of depression [33].

The significant consequences of the exposure to night-time light (NTL) with the fact that 54% of world's population lives in urban areas stimulates the interest in understanding how the light pollution evolves with the size of the US cities [35]. Bettencourt *et al.*, as introduced in the previous chapter, found the cities in the US exhibit three different types of allometric laws for urban indicators with population size [5]. From the results shown by Bettencourt *et al.*, several studies have been carried out on the allometry of urban indicators in different levels of human aggregation [7, 10, 38, 39]. Following this aim, we analyze and classify the allometric law between the NTL and the population of the US cities.

Here, we use three geo-referenced dataset: the population dataset, the NTL dataset and the Metropolitan/Consolidated Metropolitan Statistical Area (MSA/CMSA). In order to define the boundaries of each US city, we use two methods: the City Clustering Algorithm (CCA) [8, 9, 11] and the MSA/CMSA [40]. Finally, we find the allometric scaling between the NTL and the population for the two applied methods. Furthermore, to compare them, we analyze the allometric scaling between area and population.

3.2 Database and methods

Population dataset (GPWv4): The population dataset is extracted from the fourth version of the Gridded Population of the World (GPWv4) [41, 42] from the Center for International Earth Science Information Network (CIESIN) at the Columbia University. The GPWv4 models the human population distribution on a continuous surface at high resolution. Population input data is collected through several censuses around the US, between 2005 and 2014. Data are provided in grid form, where each cell is formed by 30 arc-second angles (approximately 1 km $\times$ 1 km at the Equator line). We use the US population count data, measured in number of people, for the year 2015, as depicted in Fig. 3a.

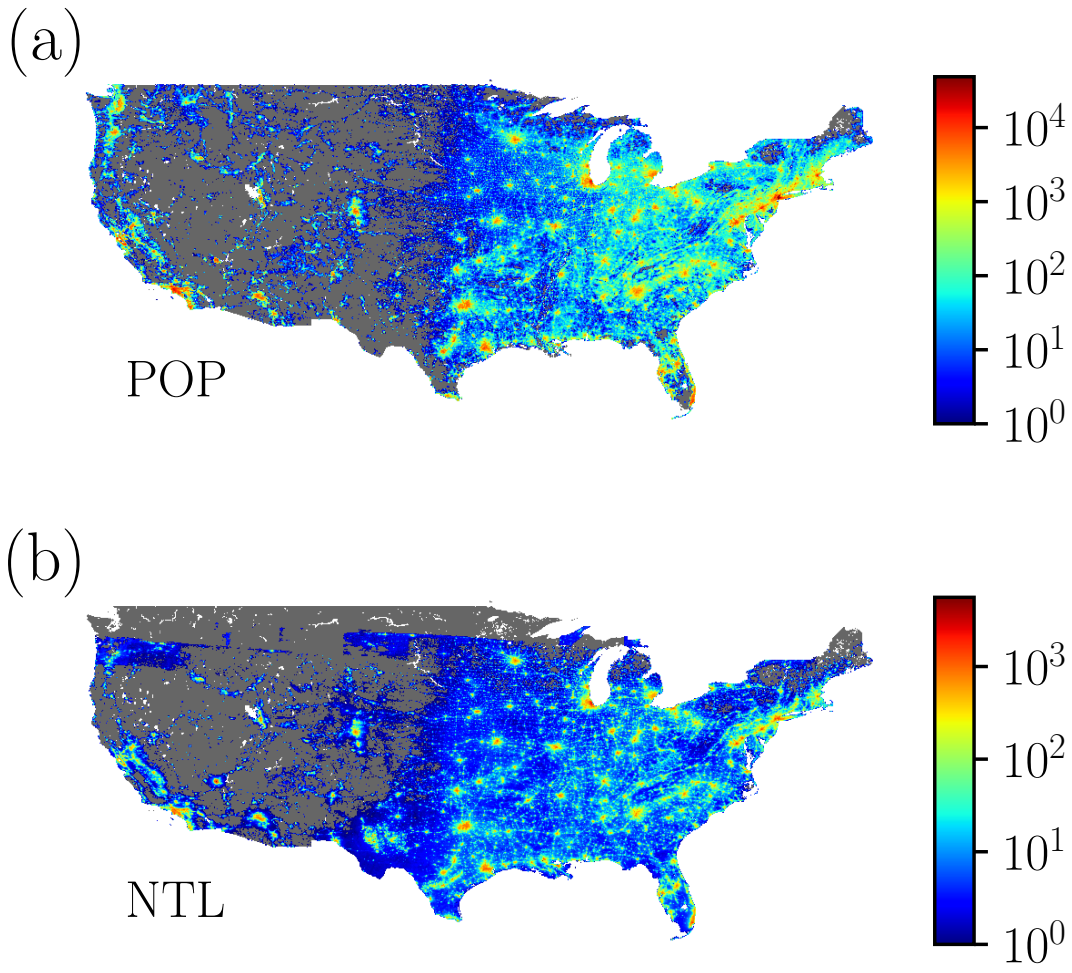

Figura 3: **Datasets (on colors).** (a) The population dataset is defined as a 30 arc-second geolocated grid. It is obtained from the GPWv4 in logarithmic scale for the year 2015 [41, 42]. (b) The NTL dataset is obtained through the night-time light radiance emission data from the VIIRS DNB in $nW/cm^2/sr$ [43–45]. It is defined at the resolution of 15 arc-second grid in logarithmic scale for the year 2015 (April).

The method successively introduced requires the population density of each grid cell. The-

refore, we calculated the area of each grid cell dividing them into two spherical triangles. The area of a spherical triangle with edges a, b and c is given by,

$$A = R^2 E, \tag{3.1}$$

where $R = 6,378.137$ km is the Earth's radius and the spherical excess E is defined by the following expression:

$$E = 4\tan^{-1}\left[\tan\left(\frac{s}{2}\right)\tan\left(\frac{s_a}{2}\right)\tan\left(\frac{s_b}{2}\right)\tan\left(\frac{s_c}{2}\right)\right]^{1/2}. \tag{3.2}$$

with $s = (a/R + b/R + c/R)/2$, $s_a = s - a/R$, $s_b = s - b/R$, and $s_c = s - c/R$. In this context, the distance between two points, i and j, on the Earth's surface is calculated by,

$$d_{ij} = R\theta, \tag{3.3}$$

with

$$\theta = \cos^{-1}[\sin(y_i)\sin(y_j) + \cos(y_i)\cos(y_j)\cos(x_j - x_i)]. \tag{3.4}$$

In this formalism, the values of x_i (x_j) and y_i (y_j) are the longitude and latitude, respectively, of the point i (j), measured in radians.

Night-time light dataset (NTL): The NTL dataset is given by the night-time light radiance emission data from the National Centers for Environmental Information (NCEI) [44]. The NTL dataset is defined by the monthly average of radiance, measured in $nW/cm^2/sr$, using the night-time data from the scanning radiometer Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) [43–45]. The VIIRS DNB data are processed and filtered in order to exclude data impacted by the lunar illumination, lightning and cloud-cover, but they are susceptible to other temporal lights, *e.g.* aurora, fires, and boats [43, 44]. Such data span through the entire globe with a resolution of 15 arc-second (approximately 500 m × 500 m at the Equator line) between the latitudes 75° North and 65° South. We use the US data for the year 2015 (April), as shown in Fig. 3b.

Metropolitan Statistical Area (MSA), Primary Metropolitan Statistical Area (PMSA) and Consolidated Metropolitan Statistical Area (CMSA): The MSA are geographic entities with high degree of socioeconomic integration and population over 50,000 people. The PMSA are quite similar to MSA, however they present population over 1,000,000 people. The CMSA are metropolitan regions defined by the agglomeration of some PMSA. They are all delineated by the Office of Management and Budget (OMB) and provided by the US Census Bureau [40].

Data processing: In order to superimpose the datasets, we perform two

processes: (i) As the NTL grid has a higher resolution than the GPWv4 grid, we sum the values of all NTL grid cells, which their geolocated centers are within the same geolocated GPWv4 grid cell. Therefore, we produce a new NTL grid with the same positioning and resolution of the GPWv4 dataset; (ii) For the MSA/CMSA case, we use the same approach of (i), even though the MSA/CMSA are complex polygons. To deal with this problem, we use the even-odd rule algorithm [46]. Thus, we define the NTL value for each MSA/CMSA.

City Clustering Algorithm (CCA): We define the boundaries of each US city by applying the CCA to the population grid [8, 9, 11]. We use the CCA explained in the previous chapter.

3.3 Results

We apply the CCA to the population grid varying D^* (in $people/km^2$), from 0 to 10000, and ℓ (in km), from 1 to 20. For all pairs of parameters, we find that it is possible to statistically correlate through power-law relations the area and the population as well as the NTL and the population of the US cities,

$$\log(\text{AREA}) = a + \alpha_{CCA} \log(\text{POP}), \tag{3.5}$$

$$\log(\text{NTL}) = b + \beta_{CCA} \log(\text{POP}). \tag{3.6}$$

The exponents α_{CCA} and β_{CCA} are obtained through Ordinary Least Square (OLS) [47] fitting to the data for different values of the parameters D^* and ℓ. The ranges of compatibility and the consistency of the CCA technique are investigated in Figs. 4a-d.

Indeed, the definition of the parameters D^* and ℓ of the CCA affects the dimension and the geometry of the cities, but from the Figs. 4c and 4d, it can be seen that it does not affect the allometric exponent β_{CCA}. Here, our starting strategy is to determine a range of parameters D^* and ℓ for which the relation between area and population is isometric [7, 9–11]. We find that for $D^* > 4000$ and $\ell = 3$ the allometric exponent α_{CCA} is between 0.93 and 0.95 and we consider this relation approximately linear. Inside this range, we analyze the result of the CCA using $D^* = 4560$ and $\ell = 3$, where the five larger cities in the US Northeast Coast naturally emerge, as depicted in Fig. 5. We believe that, the lack of an exactly linearity, also inside this range, is due to the high density of some downtowns, specifically, of the most populated urban centers of the US Northeast Coast.

For the pair of parameters, $D^* = 4560$ and $\ell = 3$, we find a allometric exponent $\alpha_{CCA} = 0.93 \pm 0.01$ (Figs. 6a) and a linear scaling between NTL and the population with

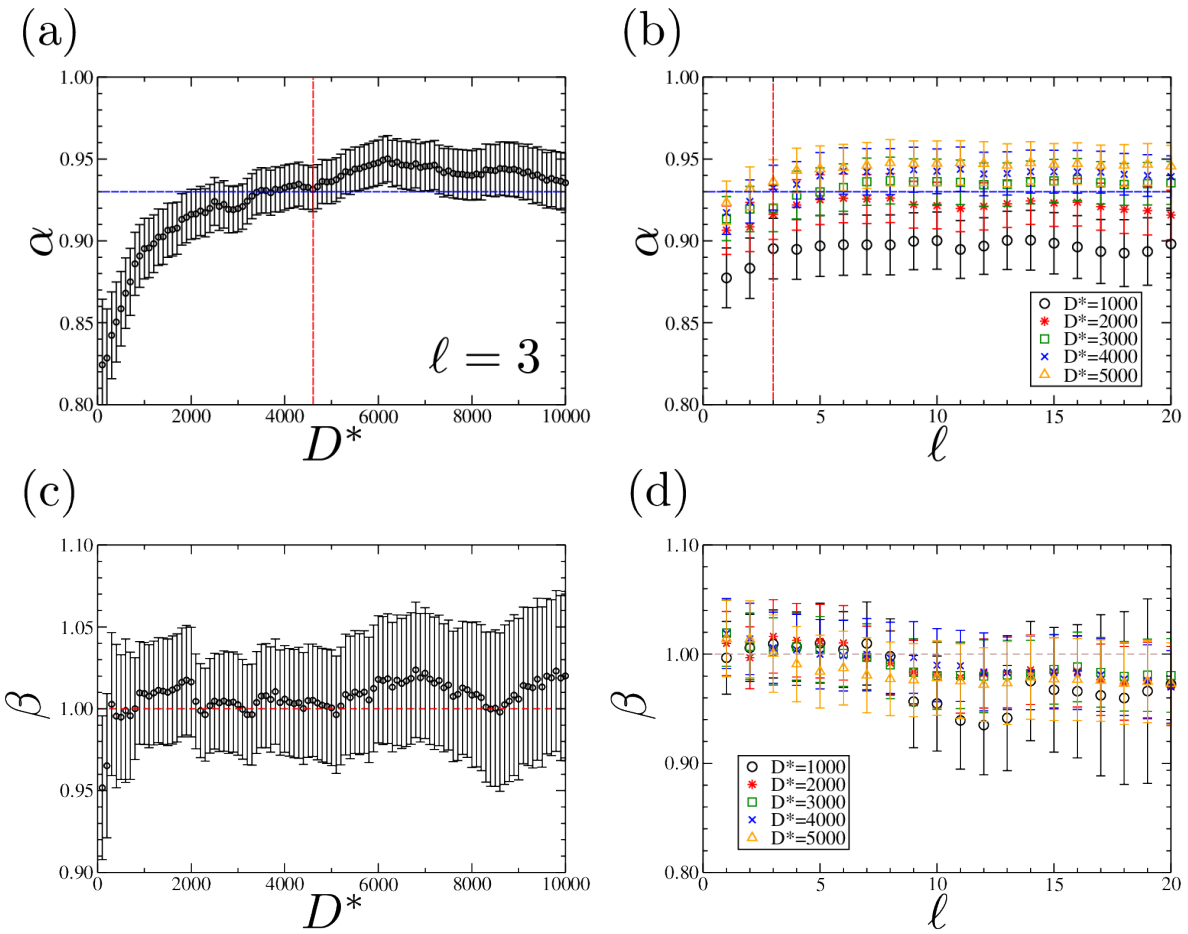

Figura 4: **Allometric exponent α_{CCA} and β_{CCA} as a function of the parameter D^* and ℓ (on colors).** (a) The exponent α_{CCA} as a function of D^* for $\ell = 3$ km. The parameter D^* varies from 0 to 10000 $people/km^2$. For $D^* > 4000$ and $\ell = 3$ the allometric exponent α_{CCA} is between 0.93 (dashed blue line) and 0.95. For $D^* = 4560$ $people/km^2$ (dashed red line) we observe the arising of five large cities in US Northeast Coast. (b) The exponent α_{CCA} as a function of the CCA parameter ℓ for D^* =1000, 2000, 3000, 4000, and 5000 $people/km^2$. We find a plateau region after $\ell = 3$ km, where $\alpha_{CCA} \approx 0.93$ (dashed blue line). (c) The figure shows the allometric exponent β_{CCA} as a function of D^* for $\ell = 3$ km. The parameter D^* varies from 0 to 10000 $people/km^2$. The dashed red line corresponds to $\beta = 1$. (d) The figure shows the allometric exponent β_{CCA} as a function of the CCA parameter ℓ for D^* =1000, 2000, 3000, 4000, and 5000 $people/km^2$. The dashed brown line corresponds to $\beta = 1$.

exponent $\beta_{CCA} = 1.01 \pm 0.02$ (Figure 7a). Alternatively, others parameters inside this range could be analyzed without affecting the allometric exponent β_{CCA} (as shown in Figs. 4c and 4d).

By analyzing the allometric scaling of the NTL with the population of the US cities using the MSA/CMSA (Fig. 7b), we obtain the allometric exponent $\beta_{MSA/CMSA} = 0.89 \pm 0.02$. Such an exponent characterizes a sublinear relation between the NTL and the population, in contrast with the CCA result.

As shown in Fig. 6b, the sublinear scaling behavior of the MSA/CMSA areas as

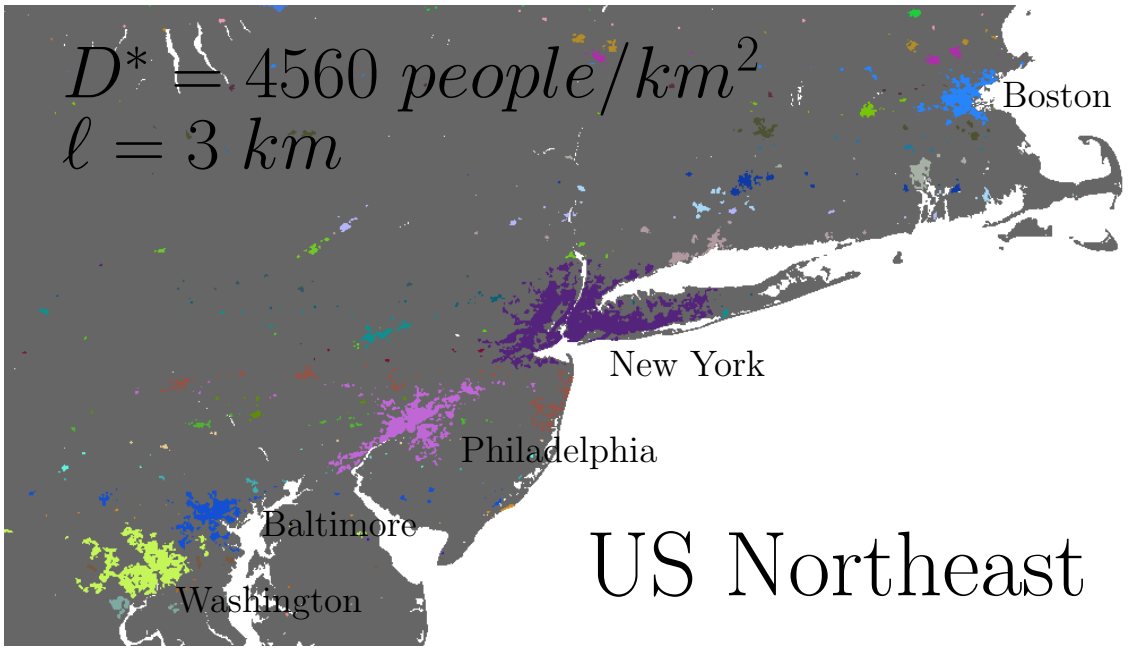

Figura 5: **Application of CCA to the US Northeast region (on colors).** We use the CCA parameters D^* =4560 $people/km^2$ and $\ell = 3\ km$. The clusters of different colors identify different urban agglomerations. Essentially, we distinguish five famous cities such as Boston (light blue), New York (purple), Philadelphia (pink), Baltimore (blue), and Washington D.C. (light green).

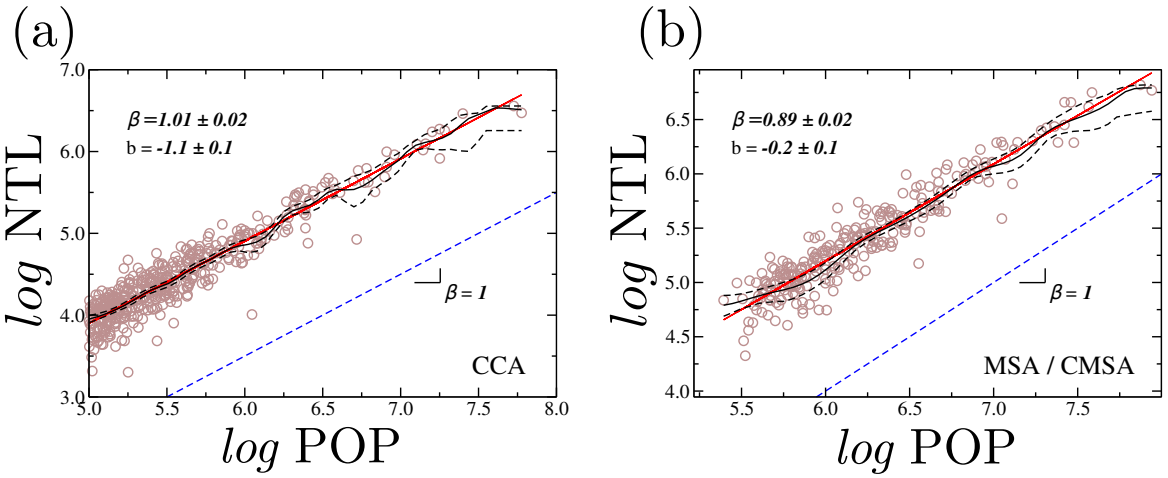

Figura 6: **Allometric exponent α applying the CCA and using the MSA/CMSA definitions (on colors).** (a) The figure shows the allometric scaling law in Eq. 3.5 and its allometric scaling exponent $\alpha_{CCA} = 0.93 \pm 0.01$ using CCA parameters D^* =4560 $people/km^2$ and $l = 3\ km$. The red line is the OLS result, and the solid black line is the N-W estimator. The dashed black lines show the 95% confidence bands of the N-W. The dashed blue line corresponds to $\alpha = 1$. (b) The figure shows the allometric scaling exponent $\alpha_{MSA/CMSA} = 0.49 \pm 0.03$ using the MSA/CMSA definitions. The red line is the OLS result, and the solid black line is the N-W estimator. The dashed black lines show the 95% confidence bands of the N-W. The dashed blue line corresponds to $\alpha = 1$.

a function of their corresponding populations, $\alpha_{MSA/CMSA} = 0.49 \pm 0.03$, clearly suggests that this might not be the most adequate definition of a city agglomerate to be adopted in our study.

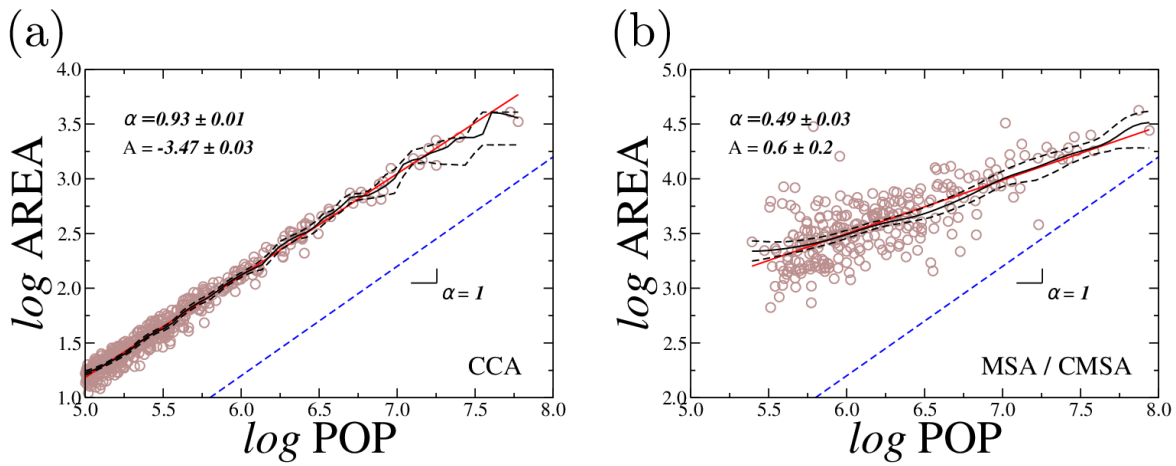

Figura 7: **NTL versus population using the CCA and the MSA/CMSA definitions (on colors).** (a) NTL versus population using CCA parameters D^* =4560 $people/km^2$ and $l = 3$ km. The graph shows a linear relation between the NTL measured in $nW/cm^2/sr$ and the population with allometric scaling exponent $\beta_{CCA} = 1.01 \pm 0.02$ ($R^2 = 0.88$). The solid red line is the linear regression obtained using the OLS method. The solid black line is the N-W estimator and the dashed black lines show the lower and the upper confidence interval (95%) [3, 48]. The dashed blue line corresponds to $\beta = 1$. (b) NTL versus population using MSA/CMSA. The graph shows a sublinear relation between the NTL, measured in $nW/cm^2/sr$, and the population with allometric scaling exponent $\beta = 0.89 \pm 0.02$ ($R^2 = 0.89$). The red line is the linear regression and the black line is the N-W estimator. The dashed black lines show the 95% confidence band of the N-W.

As indicated by Oliveira *et al.* [7], the arbitrary geopolitical concept behind the MSA/CMSA seems to overestimate the natural limits of urban areas. In order to illustrate this fact, we show in Fig. 8 the MSA/CMSA of the five most populated US regions, namely, New York-Northern New Jersey-Long Island (NY, NJ, CT, PA), Los Angeles-Riverside-Orange County (CA), Chicago-Gary-Kenosha (IL,IN,WI) and Houston-Galveston-Brazoria (TX). The first and second columns show respectively the detailed maps of the population and the NTL datasets. The third column exhibits the cities defined by the CCA with D^*=4560 and ℓ=3, as well as the discrepancy between the urban areas belonging to MSA/CMSA and CCA.

3.4 Discussion

We analyzed the allometric scaling behavior of the NTL as a function of the population of the US cities. Our results corroborate previous works showing that the

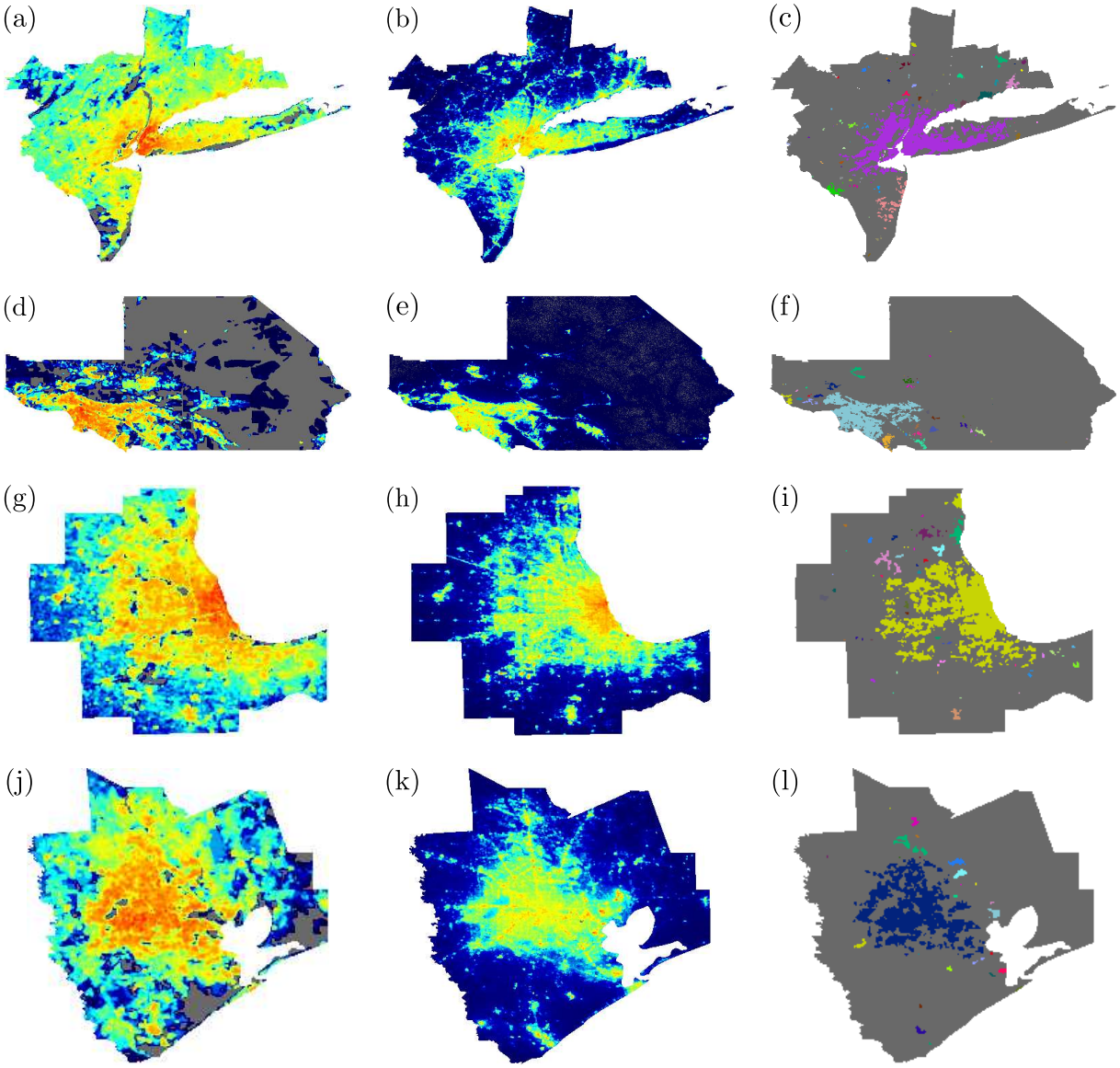

Figura 8: **Comparison between the CCA and MSA/CMSA (on colors).** Figures (a), (d), (g) and (j) are the human population grid in logarithmic scale obtained from the GPWv4 for the year 2015 [41, 42]. Figures (b), (e), (h) and (k) are the NTL measured in logarithmic scale with units $nW/cm^2/sr$ obtained through the night-time light radiance emission data from the VIIRS DNB [43–45]. In figures (c), (f), (i) and (l) we show the CCA clusters obtained using the CCA parameters D^* =4560 $people/km^2$ and $l = 3$ km of the CMSA of: New York-Northern New Jersey-Long Island (NY, NJ, CT, PA), Los Angeles-Riverside-Orange County (CA), Chicago-Gary-Kenosha (IL,IN,WI) and Houston-Galveston-Brazoria (TX). The figures show the discrepancy between the area estimated by the MSA/CMSA and the area delimited by the CCA.

scaling behaviors of urban indicators with population can be substantially different for distinct definitions of city boundaries. Precisely, using the MSA/CMSA definition, we found a sublinear allometric scaling exponent $\beta_{MSA/CMSA} = 0.89 \pm 002$. Applying the CCA, we found an exponent $\beta_{CCA} = 1.01 \pm 0.02$ which indicates an isometric relation between the light pollution and the population of the US urban agglomerations, in clear

contrast with the exponent obtained using the MSA/CMSA. Considering the consistency of the CCA definition in terms of the extensivity between land population and area of their generated clusters, as demonstrated in previous studies for other urban indicators [7], we come to the conclusion that the proportionality between light pollution and population is indeed correct, as intuitively expected [49]. Under this framework and without loss of generality, it is therefore plausible to utilize NTL as a surrogate for city population in future studies.

The isometric relation between NTL and population of the US urban agglomeration, obtained applying the CCA, imply that small and large cities are proportionally indistinguishable in terms of light pollution. In other words, there is no *economy of scale* or sublinearity concerning the NTL in US cities. Our result shows that a growth of the US cities will aggravate the light pollution and therefore the possible negative effects of the light pollution for the humans and the wildlife health.

4 DYNAMICS IN THE FITNESS-INCOME PLANE: BRAZILIAN STATES VS WORLD COUNTRIES

4.1 Introduction

Previous analysis in Economic Complexity focused in the world coutries, however large countries are often characterized by a strong internal heterogeneity. Normally they are made by richer and poorer regions. For example the GDP *per capita* (GDP_p) of the states of New York is higher than the GDP_p of Mississippi in the US [50], or the difference between Kerala and Bihar in India [51], or between the unexplored forest of Amazon and the modern state of São Paulo in Brazil [52]. While the previous literature on Economic Complexity focused on countries [14–17], in this thesis we extend the analysis to the subnational level.

The reasons to extend the analysis to subnational regions are many. First, at pure academic level, we are interested in understand if the relations and the competition among world countries are similar to the relations among entities inside a single country, and, if yes, until wich level (states or municipalities). Furthermore, a deeper knowing of the economic relations among the subnational entities inside a country could help the economists to improve the internal politics of country in order to reduce the inequalities. In this thesis, we focused on Brazil.

The *Federative Republic of Brazil* in the GDP ranking of the year 2015 is the ninth world economy [54]. Its population is 2.81% of the total World population [56] and its area (8.515.767,049 km^2), divided by its twenty-seven Federative Units [55], make it the fifth largest country of the World [56]. Its administrative organization is organized in a sequence of geopolitical structures: Union, states, Federal District and Counties. Each one is autonomous and organized according with the division of powers: legislative, executive, and judiciary. Due to the deep inequalities, but also for the good perspectives of growth, Brazil and the others Latin American countries were often a focus of economic development analysis during the last century [57–61].

In the next session, we intruduce the theoretical basic of economic complexity and its results for the world countries. Then we show our contribution for the analysis of the subnational entities. We also provide an overview of Brazil from the point of view of the Economic Complexity approach and, in this context, we compare its export basket and its Fitness with the ones of the BRIC group of countries (Brazil, Russia, India, and China) [63].

Then, we focus on the comparative study of the economies of the single Brazi-

lian states. Based on the "classical" Fitness algorithm, we introduce a new methodology, called Exogenous Fitness, able to measure the Fitness of subnational entities, and we apply it to the states of Brazil. In analogy with what was proposed in [17], we show the coevolution of GDP_p and Fitness studying the predictability of the economic growth of the Brazilian states.

Furthermore, we compare the Exogenous Fitness with: (i) the (Endogenous) Fitness -i.e, the natural application of the "classical" Fitness algorithm to the subnational entities of a country- ; (ii) the results published by the *Dataviva* platform (an application of the ECI algorithm) [64].

4.2 Database

The vast majority of data used in this project is published by *DataViva* [64]. It is an open access platform that easily allows the access to a large amount of Brazilian socioeconomic data. The database is provided by the Brazilian *Ministries*: of *Employment* (MTPS), *Development, Industry and International Trade (MDIC)* and *Education* (MEC). The project is an initiative of the *Government of the state of Minas Gerais*, *Minas Gerais Investment, Trade Promotion Agency* (INDI) and the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG) [64, 75] in collaboration with the *Sistema Mineiro de Inovaçao* (SIMI) [76], *Big Data Corp* [77] and the *MIT Media Lab* [78] . The first version was published in November 2013 and the last one, the 3.0 version, in May 2015.

The platform includes data about imports/exports products, trade partners, occupation, economic activities, basic education, higher education and universities. All data are available in several levels of aggregation: region, state, mesoregion, microregion and municipality. The crossover among data and level of aggregations allows users to access more than 1 billion visualizations.

The visualization is made through some graph types, such as: Tree Map, Stacked, Geo Map, Network, Rings, Scatter, Compare, Occugrid, Line, Box Plot and Bar Chart. Furthermore, each data and aggregations is downloadable, and easily accessible through the API architecture [79].

Here, we use the export data of each Brazilian state for the entire time interval from 2000 to 2015. Furthermore, *DataViva* provides the data of total GDP and the total population for each state for the same time interval. Combining these with the GDP deflator GDP_{defl}, published by the *World Bank* [80], we find the real GDP *per capita* of each state as:

$$GDP_p^{real} = \frac{1}{N}\frac{GDP}{GDP_{defl}}100, \tag{4.1}$$

where N is the total population of each state.

Concerning World export data, used to define the matrix M_{cp} of the World countries and to calculate the products complexity, we use data from BACI dataset [66] that is grounded on the COMTRADE dataset [67]. The database, in its extension, contains data about more than 200 countries and 5000 products classified according to a 4 digit code with categorization *Harmonized System* 2007 [81]. Data are extracted from the year 2000 to 2015. The time evolution of the GDP *per capita* of each country is published by *World Bank* [80].

4.3 Method

In this section we introduce the algorithm to quantify the development of an economy called Fitness. Furthermore, we shows all the technical tools to deeply understand the algorithm and our contribution to the calculation of the Fitness for subnational entities.

First, we introduce the Revealed Comparative Advantage, that is a quantitavive criterion that togheter with the application of a threshold and the formulation of the matrix country-products makes the index non-monetary.

4.3.1 Revealed Comparative Advantage (RCA)

The Revealed Comparative Advantage (RCA) [65] is a quantitative criterion to assess the relative advantages of a country (or of a subnational entitess) in the export of certain products compared to the average export of those products. Defining q_{cp} as the flow of the export (in US dollars) of the product p by the country c, the RCA is defined as:

$$RCA_{cp} = \frac{\dfrac{q_{cp}}{\sum_{p'} q_{cp'}}}{\dfrac{\sum_{c'} q_{c'p}}{\sum_{c'p'} q_{c'p'}}}. \tag{4.2}$$

Therefore, it is the ratio between the export of product p of a country c with respect to the export of that product in the world export.

From the calculation of the RCA for each country-product pair, we build the country-product matrix M_{cp} considering the country c an exporter of a product p only if $RCA_{cp} \geq 1$. As a consequence we set $M_{cp} = 1$. Conversaly, if $RCA_{cp} < 1$, we set $M_{cp} = 0$. In this way the matrix is binary and non-monetary.

An analogous criterion is used to define the state-products matrix M_{sp}.

4.3.2 (Endogenous) Fitness

Different studies have shown the economic relevance of the diversification of the export basket for the competitiveness of a country [14, 16]. The matrix M shows a substantial nested structure highlighted by a strong triangularity, which can be interpreted in the following way: each country approximately exports all the possible products it has the capabilities to produce [15].

In this framework, the Fitness of the country c is defined as [15]:

$$F_c = \sum_p M_{cp} Q_p, \tag{4.3}$$

where Q_p is the complexity of the product p. In this way the Fitness is proportional to the sum of its exported products weighted by their Complexity stressing the importance of having at the same time both a diversified export basket and the most complex possible products in it.

At the same time the Complexity of product p is defined as:

$$Q_p = \frac{1}{\sum_c M_{cp} \dfrac{1}{F_c}}, \tag{4.4}$$

where F_c is the Fitness of the country c. This formula is motivated by the following argument: the more the exporters of a product and the smaller their Fitness, the less its expected from the Complexity. In this manner, a state with low Fitness abruptly influences the Complexity of all the products it exports [16]. Therefore, an high Complex product is made only by few countries/states with high Fitness, while a little Complex product can be made by all the countries/states, both with high and low Fitness.

This is a system of coupled equations and there are several numericals ways to solve it. One of these is using iterations (similar to the *Google PageRank* algorithm). Therefore, the final algorithm is [15]:

$$\begin{cases} \widetilde{F}_c^{(n)} = \sum_p M_{cp} Q_p^{(n-1)} \\ \widetilde{Q}_p^{(n)} = \dfrac{1}{\sum_s M_{cp} \dfrac{1}{F_c^{(n-1)}}} \end{cases} \rightarrow \begin{cases} F_c^{(n)} = \dfrac{\widetilde{F}_c^{(n)}}{\left\langle \widetilde{F}_c^{(n)} \right\rangle_c} \\ Q_p^{(n)} = \dfrac{\widetilde{Q}_p^{(n)}}{\left\langle \widetilde{Q}_p^{(n)} \right\rangle_p}. \end{cases} \tag{4.5}$$

The elements M_{cp} are the elements of the previously discussed binary country-products

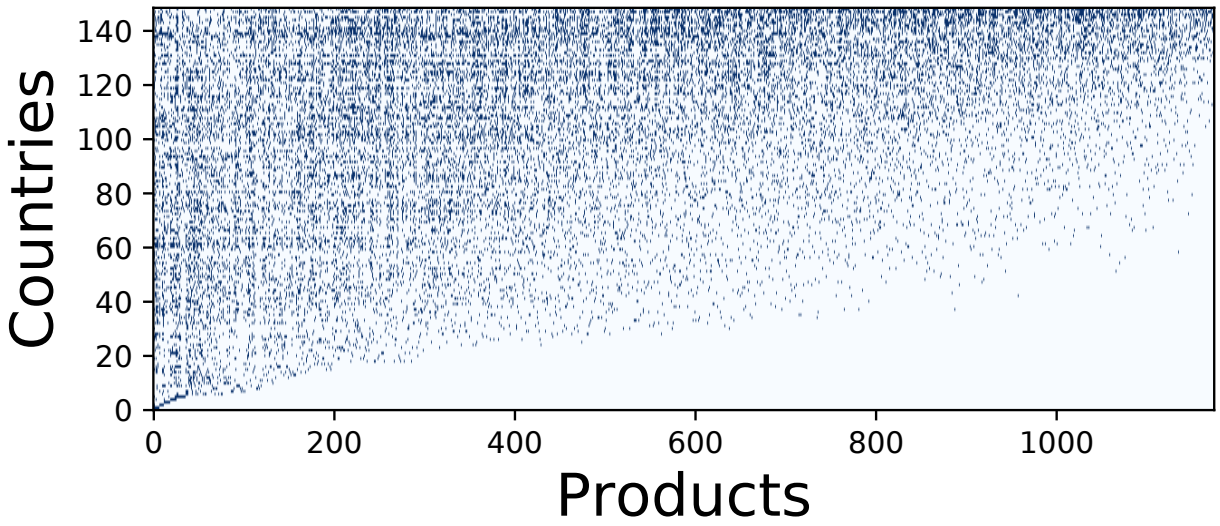

Figura 9: **The binary matrix M_{cp} of the year 2015**. The rows of the matrix represent the World countries ordered according to their Fitness with row 0 for the country with the lowest Fitness and row 147 for the one with the highest Fitness. Analogously columns represent Products ordered in terms of their Complexity from the lowest one at column 0 to the highest one at column 1174. The elements $M_{cp} = 1$ are represented as blue dots.

matrix. $\widetilde{F}_c^{(n)}$ and $\widetilde{Q}_s^{(n)}$ are intermediate variables which are subsequently normalized at each iteration. The initial conditions satisfy the relations: $\widetilde{F}_c^{(0)} = C$ and $\widetilde{Q}_p^{(0)} = C$, where we assume $C = 1$ for each country c and for each product p [16]. The stability and robustness of this algorithm has been studied in [16, 68] and the Fitness ranking of the states and the Complexity ranking of the products is unambiguously defined after the condition of convergence:

$$\sum_c |F_c^{(n)} - F_c^{(n-1)}| < \epsilon \tag{4.6}$$

Fig 9 shows the matrix M_{cp} of the World countries of year 2015 obtained by ordering the countries according to the Fitness and the products according to the Complexity. In that year, Brazil is ranked in the 44th/147 position (equivalent to the raw 103 in Fig 9).

While Fig 10 shows the matrix Brazilian states-products M_{sp} of the year 2015, by ordering the states according to the Fitness (the upper the higher complexity), and the products according to the Complexity (the more right the higher the complexity).

Furthermore, in Fig 11 we show the products *spectroscopy* [69] of the years 2005 (dashed lines) and 2015 (filled colors) for few Brazilian states such as: São Paulo, Paraná, Ceará, and Roraima. The spectroscopy is a graphic representation of the export volume (in US Dollars) of a state for each product with $M_{sp} = 1$ ordered at increasing Complexity from left to right [69]. We subsequently group the products (10 for bin) and we summed the export volumes of each product inside each bin. The spectroscopy allows

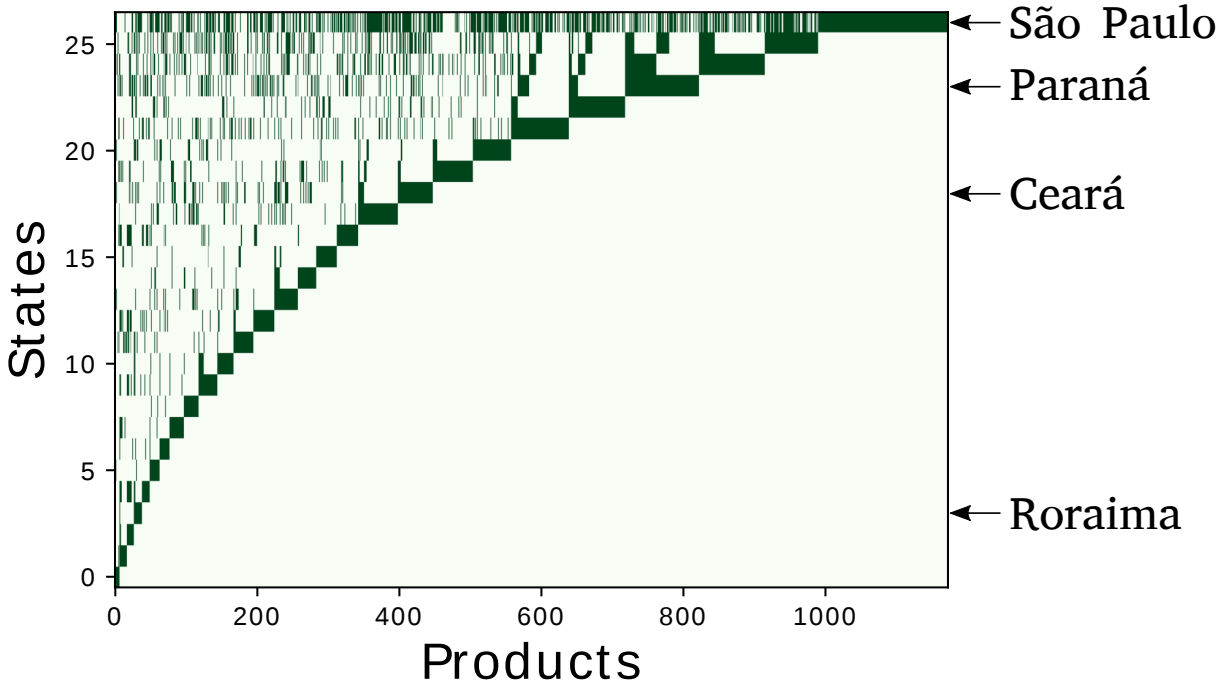

Figura 10: **The binary matrix M_{sp} of the year 2015**. Each row of the matrix represents a Brazilian state. States are ordered in terms of their Fitness from the smallest value (row 0) to the largest one (row 26). Analogously columns represent Products ordered in terms of their Complexity from the smallest value (column 0) to the largest one (column 1172). The matrix elements M_{sp} are drawn in dark green and the others in white. In the figure we highlight high Fitness states such as São Paulo and Paraná, a middle rank State such as Ceará and a low Fitness state such as Roraima.

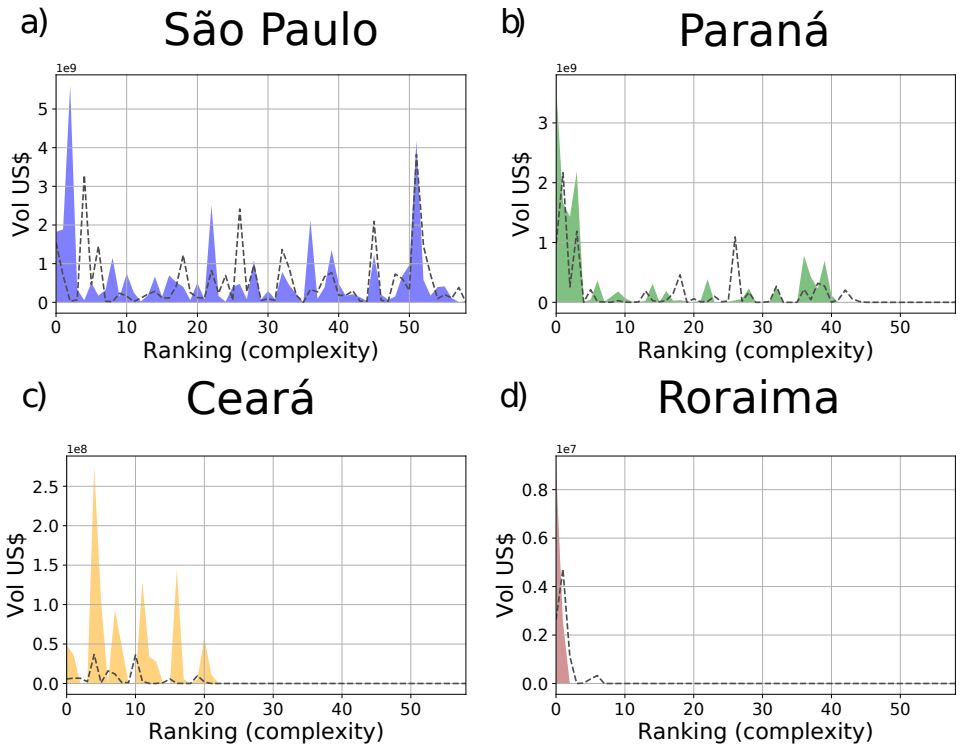

Figura 11: **Products spectroscopy of the years 2005 (dotted lines) and 2015 (filled colors) of the states:** a) São Paulo, b) Paraná, c) Ceará, and d) Roraima. The figures show the export volume (in US Dollars) of those states for each product with $M_{cp} = 1$ ordered according to their Complexity. Products are grouped in bins of 10 and the export volume in each bin are summed up.

to compare the diversification and the Complexity of the exportation of the states. The figure shows the spectroscopy of high Fitness states such as São Paulo (diversified all along the Complexity spectrum) and Paraná (with a clear peak on medium-high Complexity products), a middle rank state such as Ceará and a low Fitness state such as Roraima (with few low Complexity exports). From the figure, it emerges that a very developed state such as São Paulo has a high flow of exports for a very diversified number of products with a bias towards the high Complexity ones. Paraná has a high peak in several complex products, while Roraima has only one peak in the less complex products. Ceará is a middle ground between the two.

4.3.3 Exogenous Fitness

Here, we define the new Exogenous Fitness algorithm, an innovative method to calculate the Fitness of subnational entities of a country grounded on the measure of the products Complexity from the World-wide trade network. Exogenous Fitness is a coherent extension of the "classical" Fitness algorithm [16], with the assumption of an obvious concept: products have an intrinsic Complexity, reflected by the trade on the global World scale by all countries, while the trade from the regions of a single country

may not represent well such intrinsic Complexity as it can be affected by local biases. In particular if we consider only Brazil to define the Complexity of the exported products, we can introduce local economic biases in its measure related to the peculiar features of Brazil economy. Indeed, as shown in Fig 10, there is a big range of products made only by few states that make the measure of Complexity very inaccurate. From this observation, it is natural to use as the best measure of Complexity of products the ones Q_p^W extracted from the Fitness algorithm applied to the trade of goods of all World countries, i.e. we take:

$$Q_p^W \equiv Q_p^B \equiv Q_p. \tag{4.7}$$

Indeed, the Complexities of the products obtained applying the Endogenous Fitness to the World countries (Q_P^W) can be considered the same of the Complexities of the products inside Brazil (Q_P^B) and, therefore, we simply define them as Q_p.

Therefore, the algorithm consists of two steps:

1. We apply the (Endogenous) Fitness (eq. 4.5) to the World countries, as previously done in [15–17]. The criterion adopted to determine if a country c is a "good" exporter of a given product p is again based on the RCA extended to all World countries: we set $M_{cp} = 1$ if $RCA_{cp} \geq 1$ and $M_{cp} = 0$ otherwise (see the section *Database* for the source of the data). Applying the (Endogenous) Fitness algorithm to the matrix M_{cp}, after a sufficiently large number of iterations the algorithm converges to the fixed point so that, we obtain the respective Fitness F_c for each country and the Complexity Q_p^W for each product.

2. From the assumption eq. 4.7, we use as Complexity of the products exported by Brazilian states Q_p the values obtained by the Fitness algorithm applied to the export of all World countries. Therefore, we use the information in the matrix M_{sp} and the product Complexity Q_p to calculate the Fitness of the Brazilian states through the following formula:

$$\begin{cases} \widetilde{F}_s = \sum_p M_{sp} Q_p \\ \\ F_s = \dfrac{\widetilde{F}_s}{\left\langle \widetilde{F}_s \right\rangle_s}. \end{cases} \tag{4.8}$$

The relevance of developing the Exogenous Fitness measure is two folds. First of all, using world wide data we extract all the information to compute the Complexity of products to better compute the Fitness of states. Since the algorithm works by exploiting

differences of capabilities, using world wide data we gain additional information related to the export baskets of countries with a wider range of Fitness and capabilities. Of course we still expect the two measures to be highly correlated in rank, in particular for a country like Brazil that contains such a vast array of development levels. As we will see in section *Comparison with other techniques*, this is indeed the case. The second reason is that the Exogenous Fitness allows to have for states Fitness values comparable with those of countries. Indeed, while the ranking between Exogenous and Endogenous Fitness are highly correlated, their actual values and distributions are vastly different. As detailed explained in the paper [70], while the ranking for the Fitness measure is always well defined, the shape of the matrix directly affects the convergence properties of the algorithm to a polarized distribution. Employing the Exogenous Fitness method we have smoothly changed values that allows for the forecasting exercises of Section *Results*.

4.4 Overview of Brazil

First, we analyze Brazil as a whole applying the (Endogenous) Fitness to World countries in the time interval from 1995 to 2015.

In Fig 12, we show the dynamics of the World countries in the Fitness-Income plane emphasizing the BRIC countries (Brazil in green, Russia in blue, India in orange, and China in red). The figure shows that India and China have in 1995 lower values of GDP_p than Brazil and Russia, but higher values of Fitness. According with [17], this difference justifies the dynamics in the plane of the four countries for the next years. Indeed, India and China continued their economic growth during the following years, while Russia and Brazil entered a period of recession [71].

In order to zoom on the differences among the dynamics of the BRIC countries, we analyze the variation of the Fitness of such countries during the interval from 2003 and 2013. The variation of the Fitness can have two different causes: (i) changes in the export basket, (ii) changes in the products Complexity. We can decompose the variation of Fitness [72] as:

$$\begin{aligned}\Delta\tilde{F}_c = \tilde{F}_c(t_1) - \tilde{F}_c(t_0) = \sum_p M_{cp}(t_1)Q_p(t_1) - \sum_p M_{cp}(t_0)Q_p(t_0) = \\ = \sum_p \Delta M_{cp}\frac{Q_p(t_1)+Q_p(t_0)}{2} + \sum_p \Delta Q_p \frac{M_{cp}(t_1)+M_{cp}(t_0)}{2}.\end{aligned} \tag{4.9}$$

where we have indicated with $\Delta X = X(t_1) - X(t_0)$ for a generic quantity X. The first term in the last step of the equation is the contribution to $\Delta\tilde{F}_c$ due to the variation in the export basket, while the second one is the term due to variation of products Complexities.

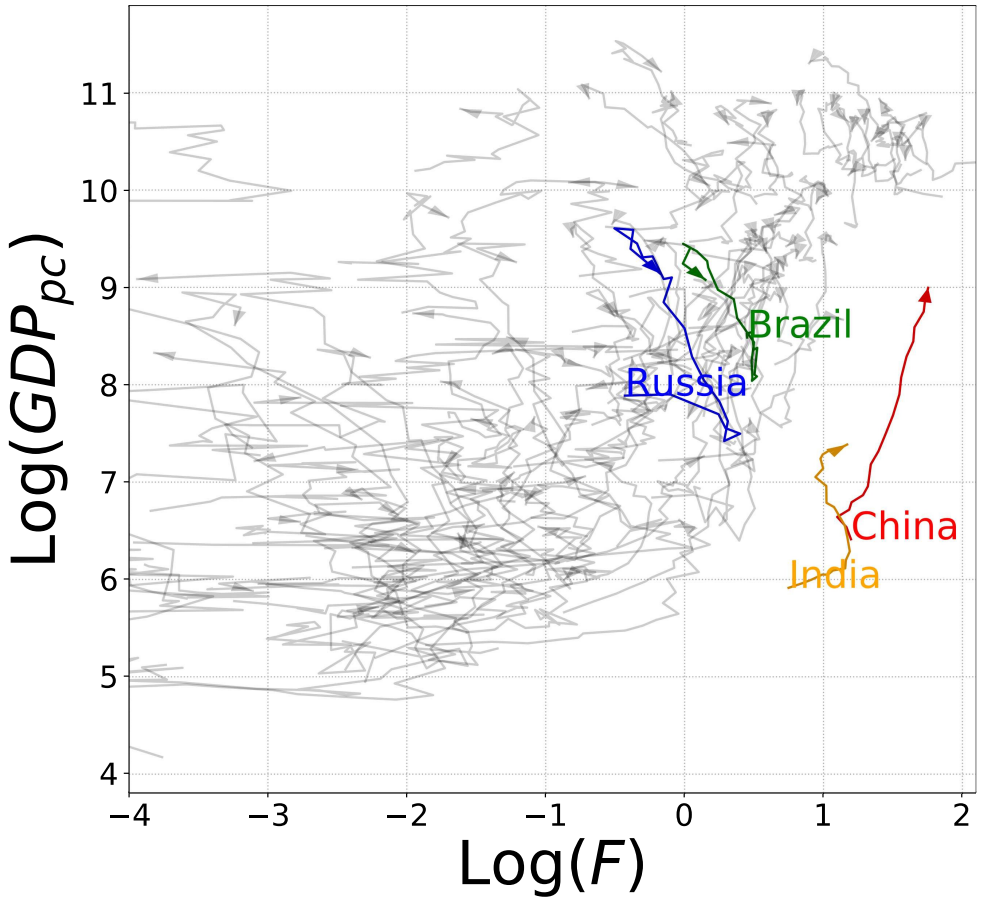

Figura 12: **Dynamics of the World countries in the Fitness-Income plane.** The figure shows the dynamics (from the year 1995 to the year 2015) of World countries in the Fitness-Income plane in logarithmic scale. We emphasize the BRIC countries: Brazil in green, Russia in blue, China in red, and India in orange.

Tabela 1: Fitness variation from 2003 to 2013 of BRIC countries.

	Variation due to changes in the export basket	Variation due to changes in the products Complexities
Brazil	-43%	-6%
Russia	-37%	-21%
China	+32%	+18%
India	-18%	+2%

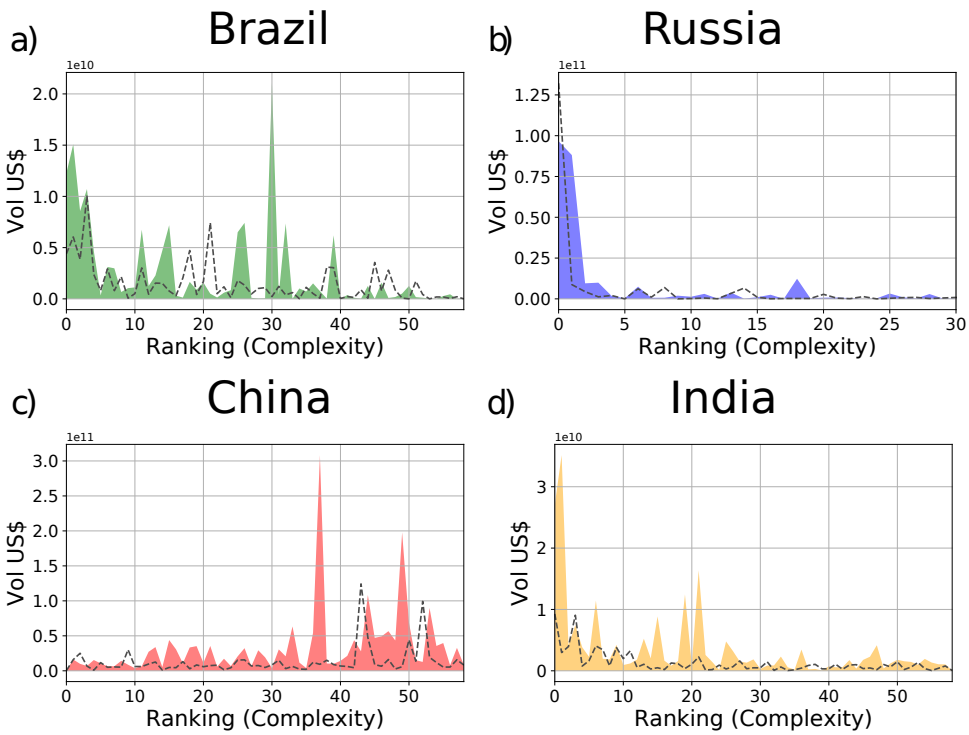

Figura 13: **Products spectroscopy of the years 2005 (dotted lines) and 2015 (filled colors) of the countries:** a) Brazil, b) Russia, c) China, and d) India. The figures show the export volume (in US Dollars) of those states for each product with $M_{cp} = 1$ ordered in terms of their Complexity. The products have been grouped (10 for bin) and the export volumes of each product inside each bin have been summed.

In Table 1, we show both the percentage variations due to the two terms. The results show a deep decrease of both terms for Russia and we can see how the loss of competitiveness of Brazil is mostly due to the drop of products that were previously exported, and not so much related to the change in complexity of those products. In contrast China has increased its export basket and the Complexity of the exported products. Instead, India in 2013 exports more complex products, but has decreased its exports diversification.

Furthermore, we show in Fig 13 the products spectroscopy [69] for the BRIC countries of the year 2005 (dotted lines) and 2015 (filled colors). The figure shows that Brazil and Russia have a high exportation only of simple products, while India and China have a high exportation of complex products.

Therefore, Figs 12-13 and Table 1, show that China and India both have a

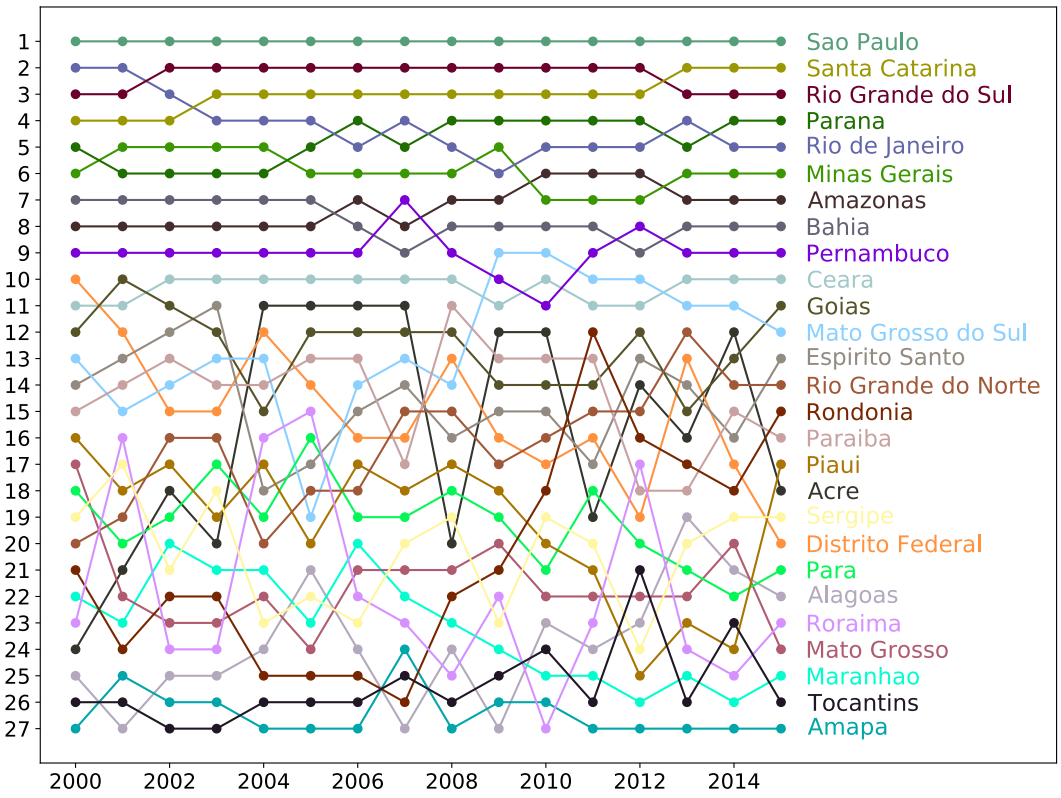

Figura 14: **Time evolution of the ranking of Brazilian states according to the Exogenous Fitness algorithm.** The figure shows the time evolution of the ranking of the Brazilian states according to the Fitness obtained through the Exogenous Fitness algorithm applied to the time interval 2000-2015.

diversified export basket and export complex products. Such factors determine a high Fitness and consequently a growth of the GDP_p in the subsequent years. On the contrary Brazil and Russia export simple products with a consequently low Fitness so that these countries entered a recession period [71].

In the next section we show the results of a deepened analysis of the internal economy of Brazil through the application of the Exogenous Fitness to the Brazilian states.

4.5 Results

We applied the Exogenous Fitness algorithm to the Brazilian states in the time interval from 2000 to 2015 obtaining for each year both well-defined values of Fitness for each Brazilian state, and the ranking of states in terms of their Fitness (shown in Fig 14).

We show in Fig 15 a map of Brazil where each state is colored according to its Fitness. From the figure, it emerges Southern states have larger Fitness, and therefore have a better economic development, than Northern states. This result is in agreement

Figura 15: **Fitness map of the Brazilian states.** The colors in the map vary from green (high Fitness) to red (low Fitness) and they show the differences of the Fitness among the Brazilian states.

Tabela 2: Fitness variation from 2003 to 2013 of the states: São Paulo, Paraná, Ceará, and Roraima.

	Variation due to changes in the export basket	Variation due to changes in the products Complexities
São Paulo	+2%	+2%
Paraná	+59%	-7%
Ceará	+53%	-10%
Roraima	-37%	+6%

with other monetary and non-monetary indices such as the Human Development Index (HDI) and the GDP [64].

Furthermore, we show in Table 2 the variation from 2003 to 2013 of the Fitness ($\Delta\tilde{F}_s$) for several states such as: São Paulo (1st in Fitness ranking of year 2013), Paraná (5th in Fitness ranking of year 2013), Ceará (10th in Fitness ranking of year 2013), and Roraima (24th in Fitness ranking of year 2013). From both Fig 11 and Table 2 we observe that São Paulo has a diversified export basket with high peaks in complex products and, at the same time, it increases both the export basket and the Complexity of the exported products in the considered time period. Paraná and Ceará, in contrast with the aggregate behavior of Brazil, in the same period grew in diversification becoming more competitive – even in the face of a minor decline in the complexity of their exported products. Roraima, on the contrary, shows a deep decrease in the diversification.

As mentioned in the previous section, Fig 12 presents the dynamics of World countries in the Fitness-GDP$_p$ plane. It shows a high degree of heterogeneity of the dynamics of countries. Indeed, the plane can roughly be divided into two regions: one with an unpredictable "chaotic" regime of the evolution of countries, and the other with a predictable "laminar" regime. In order to overcome the limitations of linear regressions, Cristelli *et al* [17] proposed an innovative data-driven non-parametric prediction scheme called the *Selective Predictability Scheme* (SPS). It is inspired by the so-called *method of analogues* [73,74] and through a *measure of concentration* it delimits predictability regions inside the Fitness-Income plane. The measure of concentration consists in dividing the plane into a grid and analyzing the time evolution of the distribution of countries inside each box with at least five countries inside.

In analogy with what has just been explained for World countries, in Fig 16a, we show the time evolution of the real GDP$_p$ as a function of the Fitness (obtained implementing the Exogenous Fitness algorithm), for each Brazilian state in the period 2000-2015. The dotted black line in the figure shows the expected level of GDP$_p$ given the level of Fitness and it is the result of the minimization of the Euclidean distance of the states from the line, weighted by the state GDP. From the figure emerges an heterogeneous dynamics similar to the dynamics of World countries that cannot be analyzed through a linear regression. Also the *measure of concentration* is not appropriate in this case. Indeed the reduced number of Brazilian states (27) compared with the number of World countries (146) makes this measure inappropriate for the internal analysis of Brazil. In order to have a significant number of cells with at least five states, the granularity of the grid should be too broad to analyze the evolution of the distribution.

Therefore, in order to validate the predictability of the dynamics of the states in the Fitness-Income plane, here we develop a novel intuitive method, the *measure of direction*. First of all let us fix the time window $[t_1, t_2]$ in which we want to study the evolution of each state in the plane $\log(Fitness) - \log(GDP_p)$. The time lag $\Delta = t_2 - t_1$ has to be taken large enough to get a sufficient noise reduction in the dynamics. We choose $t_1 = 2003$ and $t_2 = 2013$. Second, we divide the plane in a fine grid of 100×100 cells and we define two bandwidth; one for the x-axis, and the other for the y-axis. For each cell, we define around its centroid a threshold area of sides given by the two bandwidths. Then, for each cell k with at least three states at the time t_1 inside its threshold area, we computed the average dot product $\tilde{D_k}$:

$$\tilde{D}_k = \frac{2}{N(N-1)} \sum_{i<j}^{1,N} \hat{v}_i \cdot \hat{v}_j, \tag{4.10}$$

where $\hat{v}_i = \frac{\vec{v}_i}{v_i}$ where $\vec{v}_i = [log(F_i(t_2)) - log(F_i(t_1))]\hat{i} + [log(GDP_{p_i}(t_2)) - log(GDP_{p_i}(t_1))]\hat{j}$ and $\hat{i}$ and $\hat{j}$ are respectively the versors in the Fitness and GDP_p directions. N is the number of states with starting point inside the threshold area of cell k. The coefficient $\tilde{D}_k$ gives the average cosine among the versors of all states initially inside the threshold area of cell k and varies from $(-1, 1]$. It measures the dispersion of the directions of evolution in the plane in the time window $[t_1, t_2]$ of all states initially in the threshold area of cell k: when it is close to 1 all states initially in the threshold area of cell evolve in a coherent parallel way. The smaller is $\hat{D}_k$ the larger the dispersion of these trajectories. A color map of the coefficient $\tilde{D}$ in the different cells is shown in Fig 16b. From the figure it emerges that there is a region where the directions of evolution of the states tend to be parallel (showed in green) and a region where the directions of motion tend to be unevenly directed (showed in red). Increasing/decreasing the bandwidths and, therefore, the threshold area only changes the resolution of the image, but the two regions remain well-defined. In Fig 16b we used an x-axis bandwidth 0.86, and a y-axis bandwidth 0.38, providing an almost continuous variation of the colors map.

In order to investigate which is the main direction of the versors in the green region and the further directions in the red region, we divided the plane into a broader grid (10x10). For each cell we sum all the vectors inside it and then we calculate the versor of the sum vector. We show the result in Fig 16c. From one hand, from the figure we can observe a region where the states tend to evolve in the same direction (shown in green). Therefore, in this region, the future evolution of countries is predictable with good confidence. On the other hand, another region (shown in red) can be detected where the versors tend to be unevenly directed. The dynamics of the states in this region is basically unpredictable. Furthermore, in the middle of the two, there is a region of transition, shown in the figure by the overlapping of the two colors.

Lastly, in Fig 16d we show the dynamics of the states in the Fitness-Income plane highlighting in green the states with high predictability of the motion and in red those with low predictability. From the figure emerges that states as Ceará, Pernambuco, and Bahia, despite having low values of GDP, are in a region of high Predictability and, therefore, they will probably continue to growth in the same direction. While for states as Acre, Tocantins, or Alagoas the dynamics is more chaotic and predictions are less reliable.

4.6 Comparison with other techniques

In this section we compare the results obtained implementing the Exogenous Fitness with the results of the Endogenous Fitness and the ones published by Dataviva [64] obtained by applying the Economic Complexity Index (ECI).

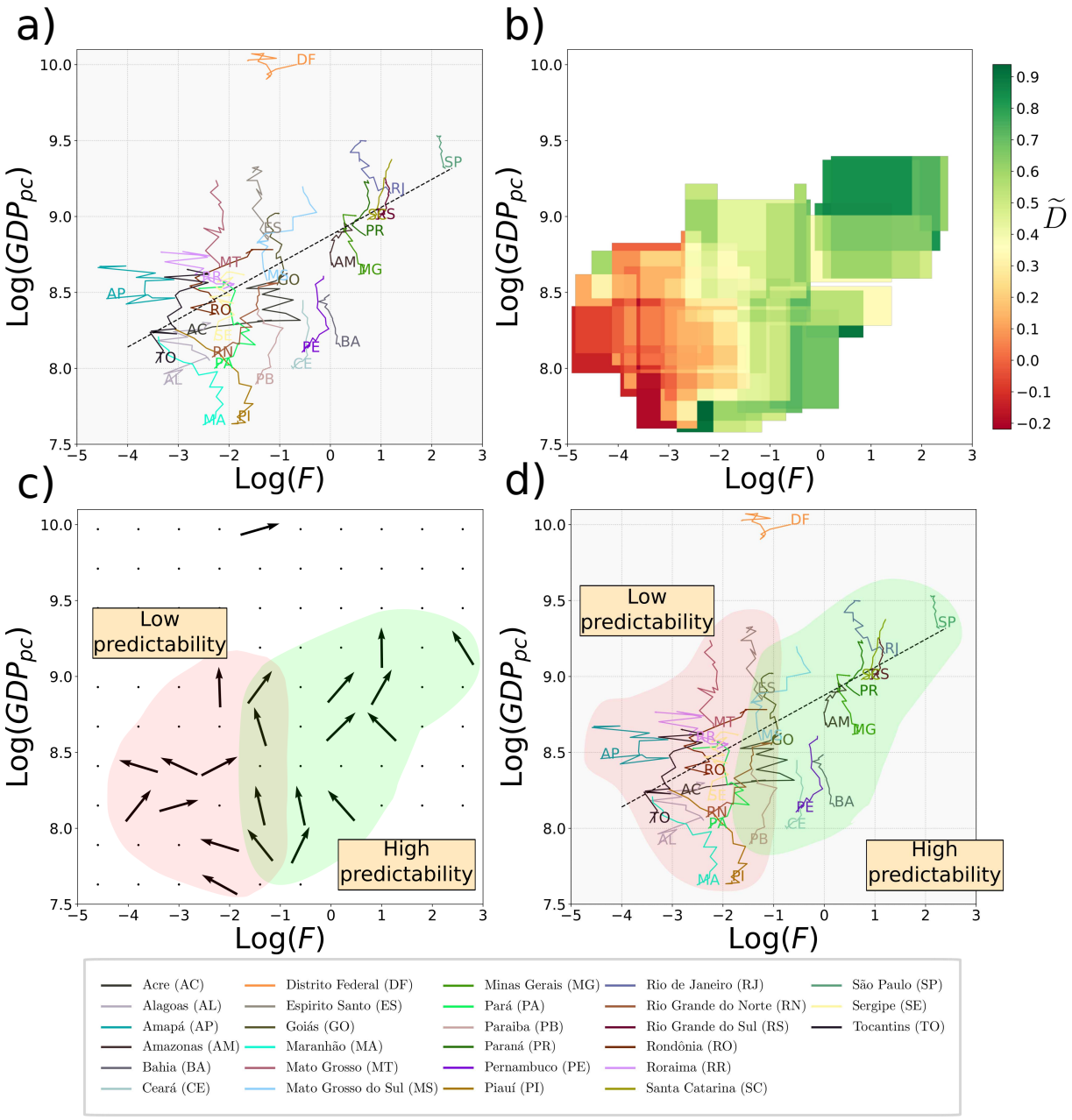

Figura 16: **Dynamics of Brazilian states in the Fitness-Income plane.** *a*) The figure shows the evolution (from 2000 to 2015) of the Brazilian states in the Fitness-Income plane in logarithmic scale. The dotted black line in the figure shows the expected level of GDP_p given the level of Fitness and it is the result of the minimization of the Euclidean distance of the states from the line, weighted by the states GDP. *b*) The figure shows the coefficient $\tilde{D}$ calculated considering a time window from 2003 to 2013. The color varies from green (where the versors of evolution tend to be parallel), to red (where the versors tend to be unevenly directed). *c*) The figure shows a grid where for each cell we calculate the versor of the sum vector. From the figure two regions appear: the first one where the versors tend to be parallel in the direction of a high GDP_p (shown in green); and the second one where the versors tend to be unevenly directed (shown in red). Figures *b* and *c* together show that there is a region (green) of high predictability of motion in direction of a high GDP_p; and a region (red) of low predictability of motion. *d*) The figure shows the dynamics (from 2000 to 2015) of the Brazilian states in the Fitness-Income plane highlighting in green the states in the high predictability region and in red the states in the low predictability one.

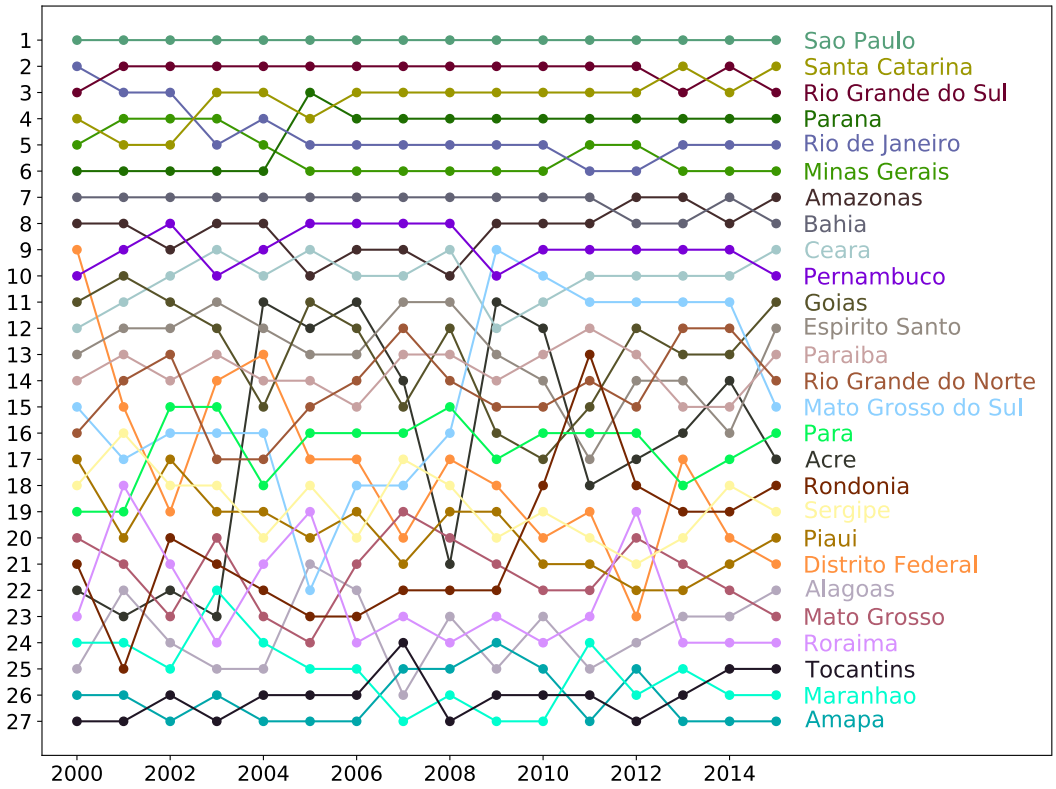

Figura 17: **Time evolution of the ranking of Brazilian states according to the (Endogenous) Fitness algorithm.** The figure shows the time evolution of the ranking of the Brazilian states in terms of the Fitness obtained through the (Endogenous) Fitness algorithm applied during the time interval 2000-2015.

4.6.1 Exogenous Fitness and Endogenous Fitness

We apply the (Endogenous) Fitness algorithm to the Brazilian states in the time interval from 2000 to 2015 obtaining the time evolution of the ranking of the states according to such kind of Fitness (shown in Fig 17). Calculating the Spearman correlation coefficient between the ranking obtained through the Exogenous and the Endogenous Fitness for each year in the analyzed time interval, we obtain an average value $\tilde{\rho}_{ExEn} = 0.97$. This result shows a strong correlation between the rankings obtained through the two different Fitness algorithms.

The Endogenous Fitness algorithm provide us a well-defined annual ranking of the Brazilian states, but not well-defined quantitative values of Fitness and products Complexity. In fact, all Fitness values except one tend to zero. After a fairly high number of iterations, however, the ranking of states stabilizes, and there are no more changes of ranking among the states. This circumstance is already been studied [70] and it is due to the shape of the matrices M_{sp}. Indeed the *external area* (where $M_{sp} = 0$) is greater than the *internal area* (where almost all elements $M_{sp} = 1$) for each analyzed year.

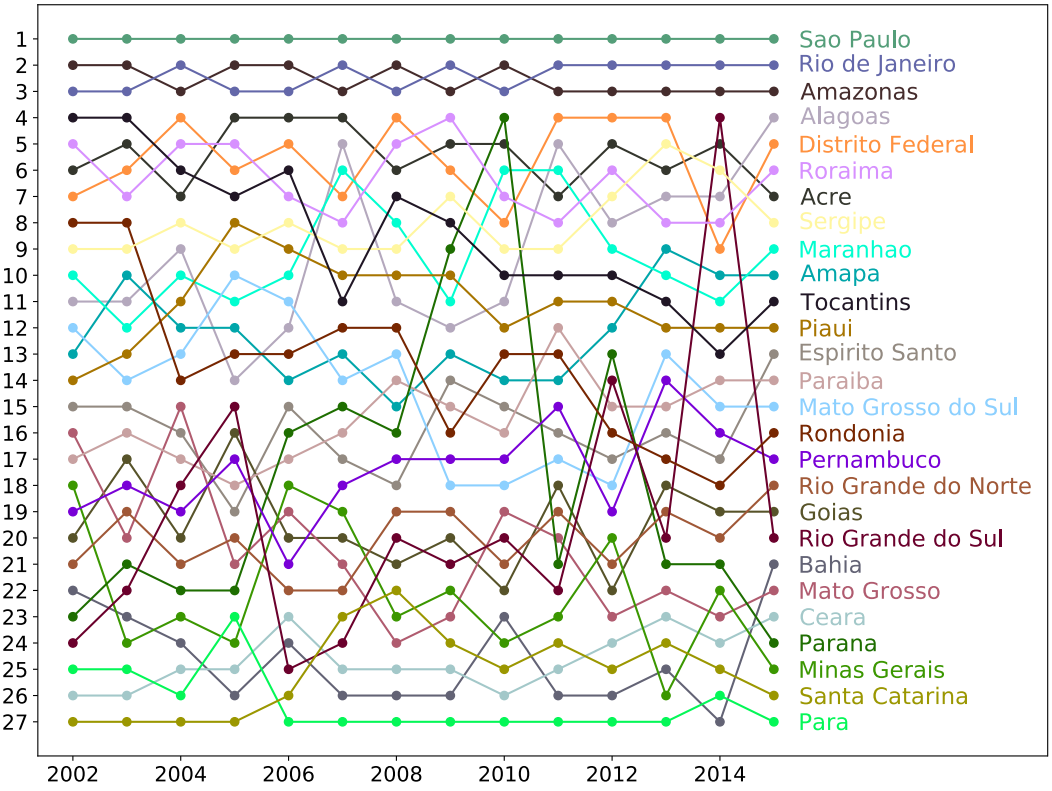

Figura 18: **Time evolution of the ranking of Brazilian states according to the ECI algorithm.** The figure shows the time evolution of the ranking of the Brazilian states during the period 2002-2015 in terms of the ECI, directly downloaded by the Dataviva platform [64].

4.6.2 Exogenous Fitness and ECI

In Fig 18 we show the time evolution (from 2002 to 2015) of the ranking of the Brazilian states according to ECI, directly downloaded by the Dataviva platform [64]. Therefore, in order to compare the ranking obtained through the Exogenous Fitness algorithm and the ECI algorithm, we calculate the annual Spearman correlation coefficient between the two rankings in the period 2002-2015, obtaining an average value $\tilde{\rho}_{ExECI} = -0.14$. This result shows an almost total absence of correlations between the two rankings, i.e. between the two algorithms.

Indeed, already from a qualitative point of view, ECI ranking seems to be unrealistic. For example, it ranks rich states in GDP, but also with high HDI [64], such as Santa Catarina or Paraná, in the last positions (respectively 26th and 24th position in 2015). Moreover, the state of Alagoas (last in HDI ranking of 2014 [64]) is unrealistically ranked in 4th position in the 2015.

In Fig 19, we show the map of Brazil where each state is colored according to its ECI. From the figure, it emerges that there is no geographic coherence among the ECI

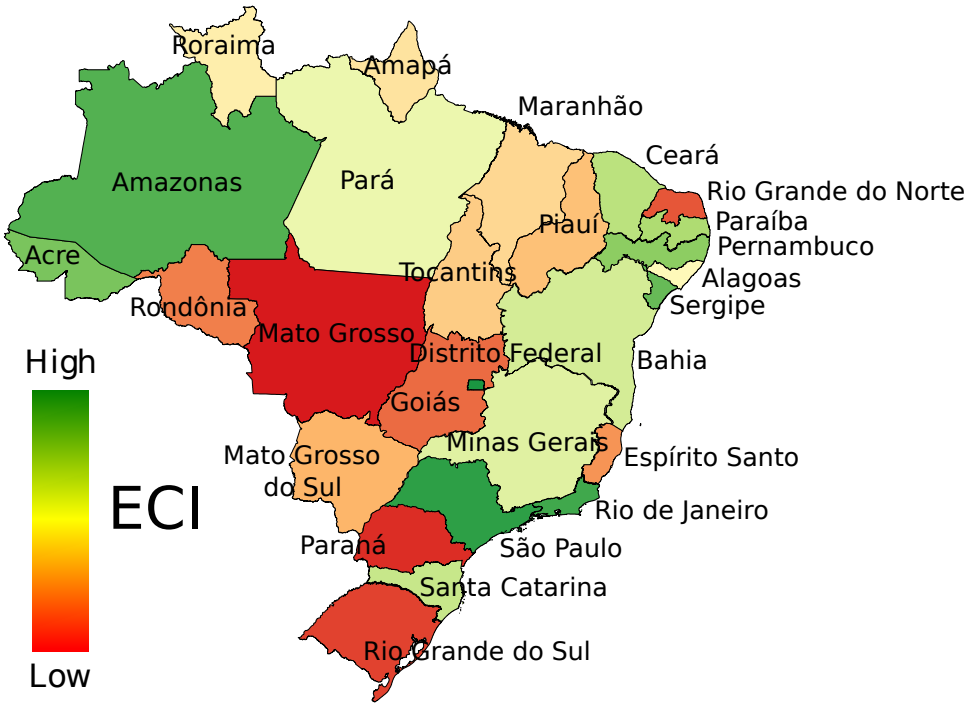

Figura 19: **ECI map of the Brazilian states.** The colors in the map vary from green (high ECI) to red (low ECI) and they show the variation of the ECI across the Brazilian states.

of the different states. For instance the figure shows that the state of Santa Catarina has a high ECI, but it is in the middle between the states of Rio Grande do Sul and Paraná that have a low ECI.

Furthermore, we show in Fig 20a the evolution of Brazilian states in the ECI-Income plane, where the income is in logarithmic scale. In Fig 20b, we show the coefficient $\tilde{D}$ above defined but applied to ECI instead to $log(Fitness)$ and in Fig 20c the directions of motions. Differently from the results obtained through the application of the Exogenous Fitness (Fig 16), using the ECI index the dynamics of the states is unpredictable. Indeed, all the states except São Paulo and the Distrito Federal are concentrated in a small region of the plane and, therefore, totally indistinguishable.

4.7 Discussion

In this project we first compared the dynamics of Brazil in the Fitness-Income plane with the other BRIC countries. In Fig 12, we observed that IC (India and China) countries, both with a high Fitness compared to the BR (Brazil and Russia) countries, grow in GDP_p for the entire analyzed time interval. Table 1 shows that IC improve the Complexity of export baskets in the analyzed time interval, and China even shows an improvement of the diversification. Instead, BR countries did not invested in diversifica-

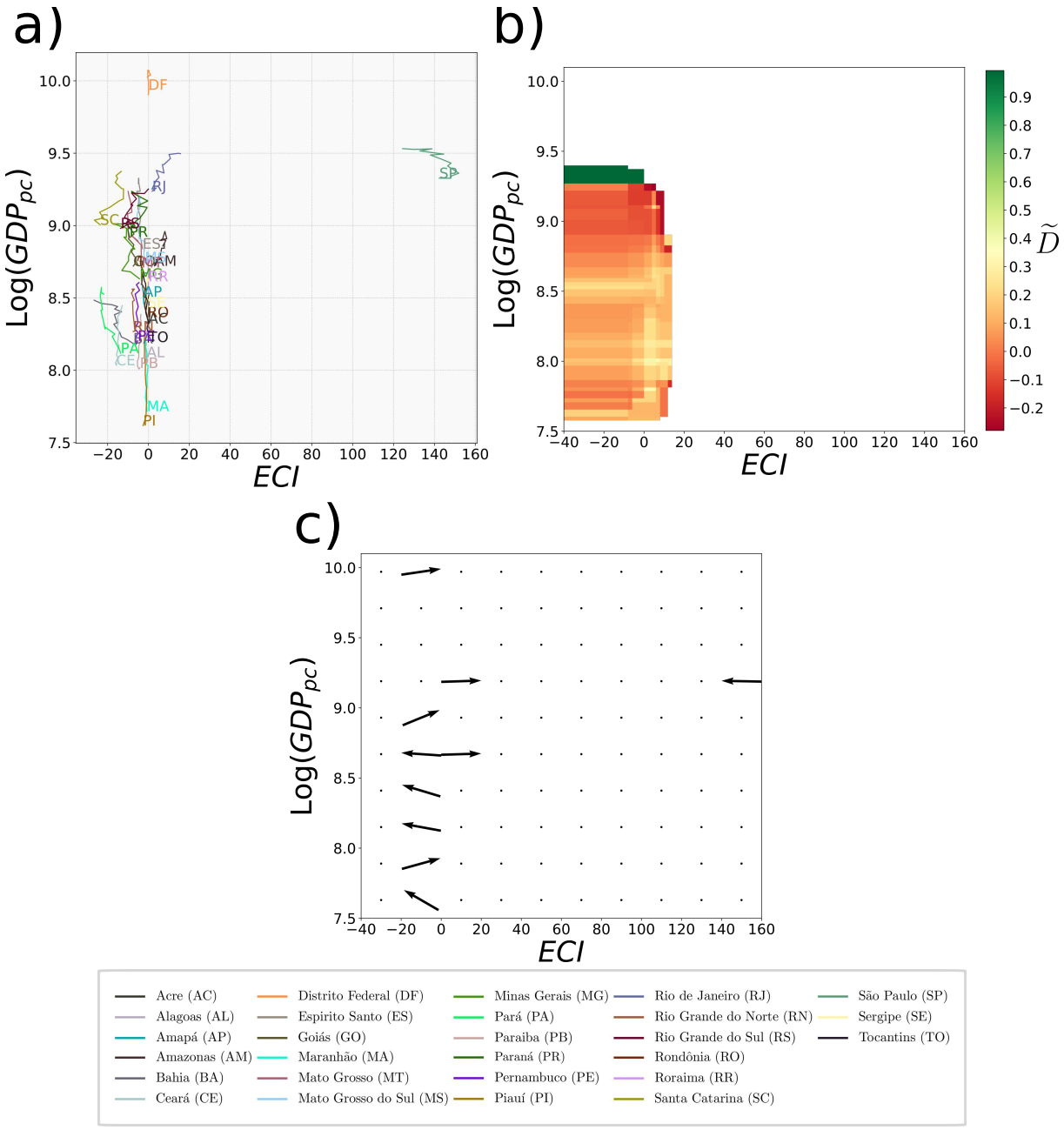

Figura 20: **Evolution of Brazilian states in the ECI-Income plane.** *a*) The figure shows the dynamics (from 2002 to 2015) of the Brazilian states in the ECI-Income plane, where the GDP_p is in logarithmic scale. Only the state of São Paulo and the Distrito Federal appear to be clearly distinguishable from the rest of the states. All the others states are indeed concentrated in a small region of the graph. *b*) The figure shows the coefficient $\tilde{D}$ calculated considering the time interval 2003-2013. Colors vary from green (where the versors tend to be parallel), to red (where the versors tend to be unevenly directed). From the figure we can therefore verify that there is a low predictability of the evolution of all the states. *c*) Here we show a grid where for each cell we calculate the versor of the sum vector. From the figure we see that there is no privileged direction, indeed the vectors are unevenly directed.

tion and in Complexity of the exported products (as shown in Table 1). These results strengthen an hypothesis previously formulated in [17]: Fitness is the driving force behind growth.

In the second part of the project, we introduced a new algorithm called "Exogenous Fitness" to calculate the Fitness of subnational entities and we applied it to the states of Brazil. The comparison between the Fitness and the GDP_p showed an heterogeneous dynamics of the Brazilian states in the Fitness-Income plane. Indeed, two regions are distinguishable in the plane: one with high predictability and the other with low predictability. Here, we have shown that economic forecasting is possible for those states in the high predictability region, while it is not for those in the low predictability region. As a consequence of this analysis Fitness seems to be the driving force behind growth. Indeed, the dynamics in the high predictability region is characterized by high values of Fitness, while high value of GDP_p is not a good signature of growth. The heterogeneous dynamics observed for the Brazilian states shows a strict analogy with the heterogeneous dynamics observed for the World countries [17]. Furthermore, by comparing the export "spectroscopy" of BRIC countries with the one of Brazilian states of São Paulo, Paraná, Ceará, and Roraima, and, comparing the variations of the Fitness, we observe that countries/states with diversified export baskets produce high complex products and grew in GDP_p in the considered period. This observation can be important for the evaluation of perspectives of economic growth for Brazilian states, and, more generally, for developing countries.

The time evolution of the ranking obtained through the Exogenous Fitness algorithm shows that developed states in the top part of the ranking change little their positions, with a smooth slow motion. On the contrary states in the inferior part of the ranking changes drastically their position during the analyzed time-interval. These facts are probably due to the stability of the developed states that are in the high predictability region of the Fitness-GDP_p plane and the instability of the states in the low predictability region.

Finally, we showed the non-correlation ($\tilde{\rho}_{ExECI} = -0.14$) between the ranking obtained though the Exogenous Fitness algorithm and the results of the ECI published by Dataviva [64]. Analyzing qualitatively the ranking of the states according to ECI, we argued that this ranking appears quite unrealistic. Therefore, we propose here the Exogenous Fitness algorithm as its valid substitute. Instead, comparing the Exogenous and (Endogenous) Fitness we obtained a strong correlation ($\tilde{\rho}_{ExEn} = 0.97$) for what concerns the ranking of states. This result shows that the two algorithmic tools are almost similar in identifying the ranking of the states, but just the Exogenous Fitness algorithm

provides also stable quantitative values of the Fitness, in addition to the ranking.

5 DYNAMICS OF RACIAL SEGREGATION AND GENTRIFICATION IN NEW YORK CITY

5.1 Introduction

Although it is not a recent phenomenon, racial residential segregation (RRS) continues to permeate the United States metropolitan areas and it is still an object of study for scientists of different areas [83–107]. The decrease of RRS in American cities is controversial and drastically varies from one city to another. Furthermore, it shows different trends according to the race analyzed. For example, several studies show that the segregation between white and black citizens has decreased in the last fifty years [91–94]. Instead, segregation between white and Hispanic, and white and Asian citizens has increased [93, 94].

Several indexes were developed to quantify RRS [83, 95–103]. The first and still most used nowadays is the dissimilarity index created by Duncan and Duncan in 1955 [103]. Subsequently, in 1988, Massey and Denton [101] defined five distinct axes of measurement of residential segregation: evenness, exposure, concentration, centralization, and clustering. The authors affirmed that, in order to fully analyze residential segregation, at least five indexes corresponding to the five spatial dimensions are necessary. Meanwhile, in 2004, Reardon and O'Sullivan's developed several measures of multigroup segregation and, among them, the authors consider the Information Theory Index the most conceptually and mathematically satisfactory measure to quantify residential segregation [99].

RRS is the cause and effect of several inequalities. Studies show the relations between racial segregation and income inequalities [104] and property values inequalities. Furthermore, RRS causes racial disparities in health and in education [104–107]. In New York City, for instance the mortality rates of black citizens vary substantially by locality according to the pattern of racial segregation [107].

In the recent years, some researches also suggest that the phenomena of gentrification is a cause of perpetuation or even of the increase of RRS [108–111]. Gentrification is defined by *The Encyclopedia of Housing* [112, 113] as:

> *The process by which central urban neighborhoods that have undergone disinvestment and economic decline experience a reversal, reinvestment, and the in-migration of a relatively well-off, middle and upper middle-class population.*

The main reason to indicate gentrification as a cause of perpetuation of racial segregation is the presumed displacement of the low-income class, in many cases predominantly black

or Hispanic citizens, from their native neighborhood during the gentrification process [108, 111, 112, 114, 115]. Taking the example of New York City once again, there is an intense debate about the gentrification of regions inside the neighborhoods of Harlem and the borough of Brooklyn [116–118].

The aim of this project is to study the dynamics of RRS in New York City from 1990 to 2010. Here, we developed a novel method able both to measure RRS and to delimit the segregated zones. Indeed, differently from previous measures, our method, in addition to quantifying the phenomena, provides a topography of the segregation. Furthermore, in the section *Comparison with the Dissimilarity index*, we compare our segregation index, the Overlap coefficient, with the dissimilarity index.

With the limit of the segregated zones, we analyze the per capita income in each high-density zone of population (defined for each race) and also in the zones of overlaps between them. In order to quantify income inequality, we calculated the Gini coefficient in each zone. Then, we studied the variation of the per capita income and of the properties' value for the census tracts that change zone during these twenty years. Finally, we focused on the segregation between white and black citizens. Particularly, we used a simplified version of the City Clustering Algorithm (CCA) [5–12] to cluster the high-density zone of black citizens and to measure the displacement and the area of the four biggest clusters. Where one of these clusters includes a gentrified region in the neighborhood of Harlem and another one is inside the borough of Brooklyn.

The project is structured as follows: first, we introduce the data and our method. Then, we present the results of the application of the method to New York City. Finally, we draw the conclusion about the results.

5.2 Database

All the data used in this project is extracted from the National Historical Geographic Information System (NHGIS) [121]. The platform provides population, housing, agricultural, and economic data with GIS-compatible boundary files for geographic units in the United States from 1790 to the present. From the platform, population data has been extracted according to race, per capita income data, and the number of owner-occupied housing units by value.

Population dataset (TABLE CW7 Persons by Hispanics or Latino origin by race): The data provides the number of people for each race for the years of 1990, 2000, and 2010 divided by *Hispanic or Latino* and *Not Hispanic or Latino*. We consider white as *Not Hispanic or Latino: white (single race)*, black as *Not Hispanic or Latino: black or African American (single race)*, Asian as *Not Hispanic or Latino:*

Asian or Pacific Islander (single race), and Hispanic as *Hispanic or Latino: white (single race)* plus *Hispanic or Latino: black or African American (single race)* plus *Hispanic or Latino: Asian and Pacific Islander (single race)*. The data table is downloadable with the respective GIS-compatible boundary file formed by census tracts standardized to the 2010 census [122].

Per capita income dataset (BD5 Per capita Income in the Previous Year): The data provides the average per capita income of each American census Tract in the previous year of 1980, 1990, 2000, and between 2008 and 2012. The values are not adjusted for inflation.

Properties values dataset (NH23 Specified owner-occupied housing units and B25075 Owner-occupied housing units): The properties values data are divided into two databases: the table NH23, for the year of 1990, and the table B25075, for the years between 2006 and 2010. The tables provide the number of houses in each price range. The price ranges are divided as: in the table NH23, in twenty ranges, and, in table B25075, in twenty-four ranges from zero Dollar to infinity. For each tract, the weighted arithmetic mean of the properties values has been calculated. The table B25075 is provided in the 2012 census tract and it is consistent with the Population data and the per capita income data, whereas table NH23 is provided in 1990 tracts. Therefore, through a superimposing process, the data was recomposed in the 2012 Census Tract. The superimposing process consists in considering all the properties in a 1990 census tract with centroid in a 2012 census tract as part of that 2012 census tract.

5.3 Method

The method consists of the following steps: first we define the limits of the city using the City Clustering Algorithm (CCA) [5–12]. Second we find the high-density zones for white, black, Asian, and Hispanic citizens. Finally, we measure the RRS through the Overlap Coefficient.

The CCA is an algorithm introduced to define boundaries of metropolitan areas [5–12]. Its result depends on two parameters: a population density threshold D^* (in $people/km^2$), and a cutoff length ℓ (in km). The elementary information for population data are provided in *census tract*. Where the tracts are geographic regions defined by the United States Census Bureau [122]. For each tract, we have the total area and the total population given by the sum of people of each race. Therefore, for each tract, its population density is calculated. According to the CCA, the assumption is that only the tracts with $D_i > D^*$ are populated.

The next step of the algorithm is the clusterization. In this step, we define the

urban center. For each populated tract, we draw a circle of radius ℓ with center in the centroid of the tract. All populated tracts that have the centroid inside the circle belong to the same cluster, and, therefore, the same city. The parameter D^* and ℓ are chosen respecting the isometry between area and population of the cities [6,7,10]. The algorithm is applied in the entire country and, subsequently, we extract only the cluster equivalent to New York City.

The importance of using the CCA to define the urban area of New York City is due to the fact that RRS deeply depends on the definition of urban areas [91, 95, 97]. For example, it was shown in [6, 7] that the Metropolitan Urban Areas (MSA) have large inhabited regions. Instead, the aim of our research is to analyze RRS in a very dense urban area, specifically in New York City.

We define the high-density (HD) zones as regions inside the city with a high population density of a specific race. The HD zone of a specific race r is defined applying a density threshold D_r^* and considering populated with that race only the tracts with $D_r > D_r^*$. D_r is the population density of that race. The choice of parameter D_r^* is made by studying how the fraction of population of race r, with respect of the total population of the same race inside the whole city, depends on it. Therefore, for each race r, we define a parameter p_r as:

$$p_r = \frac{Population\ of\ race\ r\ inside\ the\ HD\ r\ zone}{Total\ population\ of\ race\ r\ inside\ the\ city}. \tag{5.1}$$

To make the analysis as uniform as possible, we choose D_r^* so that both D_r^* and p_r take similar values for all considered races r.

In the Fig 21 we show the variation of the parameter p in function of parameter D_r^* for each race in New York City. We consider the same fraction of people in three cases using a similar D_r^*: when it is next to 0, to ∞, and ~ 2000. The first two are trivial, in fact they shows respectively all and any population. While in $p = 0.8$ (that is considering the 80% of the total population for each race), we have for each race $D^* \sim 2000$. The dotted black line in the Figure is exactly in $p = 0.8$ showing the 80% of the total population of each race. Parameter p_r has been tested in the interval from 0.7 to 0.9 without find deep discrepancies in the results. Therefore, at the end of this step, the method provides well-defined geographic limits of the HD zones for each race.

From the definition of the HD zones, we measured the RRS between two races computing the sharing area (or overlap area) between the two HD zones. Therefore, we define the Overlap coefficient (or Szymkiewicz-Simpson coefficient [119]) as:

$$O_{rr'} = \frac{|X_r \cap X_{r'}|}{min(|X_r|, |X_{r'}|)}, \tag{5.2}$$

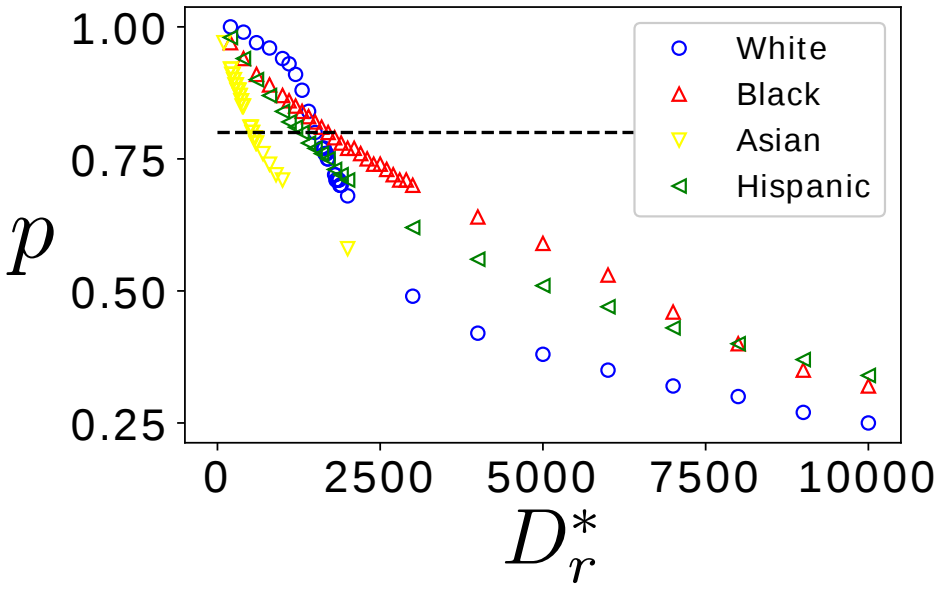

Figura 21: **Variation of p in function of parameter D_r^* for each race in New York City in 2010.** The Figure shows the variation of parameter p in function of parameter D^* for white, black, Asian, and Hispanic. The dashed black line in $p = 0.8$ shows the 80% of the total population for each race.

where X_r and $X_{r'}$ are respectively the HD zone areas of races X_r and $X_{r'}$. Coefficient $O_{rr'}$ is the sharing area between the HD r zone and the HD r' zone divided by minimum area between the two zones. The Overlap coefficient is included between 0 and 1. When it is next to 0 (low overlap), the coefficient indicates high segregation, while when it is next to 1 (high overlap), it indicates low segregation (see Table 3).

5.4 Results

Firstly, we defined the limits of New York City by applying the CCA to the population data in 2010. Then, we calculated the HD zone for white, black, Asian, and Hispanic for the year of 1990, 2000, and 2010. In Fig 22, we show the dynamics of the segregation between: white and black; white and Hispanic; and white and Asian citizens with the respective Overlap zones.

For each pair of races, we calculated the Overlap coefficients and the results were presented in Table 3. The Table shows that the segregation between white and black, and black and Asian citizens remain quite stable during the time interval. While segregation between white and Hispanic, white and Asian, and Hispanic and Asian has increased, the segregation between black and Hispanic citizens has decreased. Black people are constantly the most segregated having a high overlap coefficient only with Hispanic.

After the definition of the HD zones and the Overlap zones, we calculated the

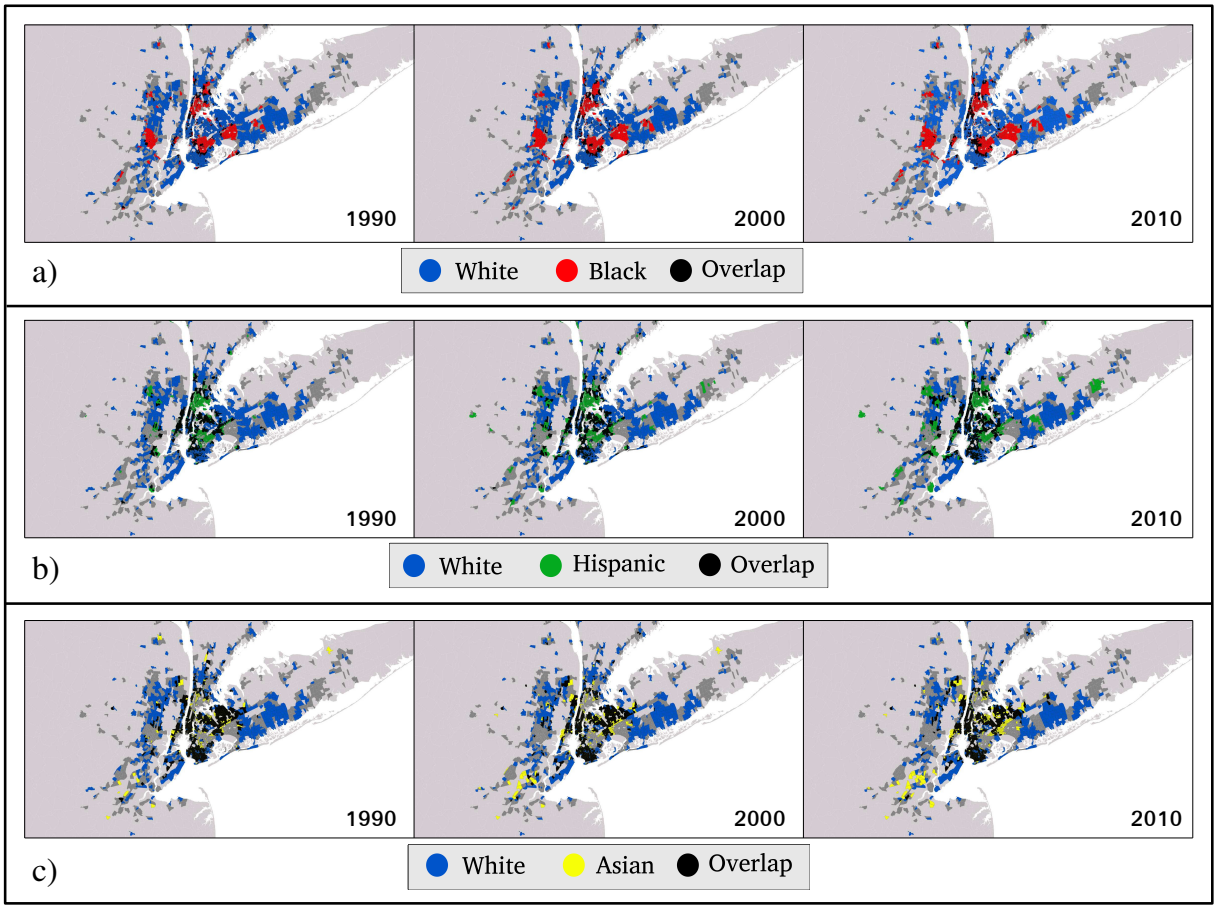

Figura 22: **Dynamics of the HD zones of:** a) white (in blue), black (in red), and Overlap between white and blacks (in black). b) whites (in blue), Hispanics (in green), and Overlap between whites and Hispanics (in black). C) whites (in blue), Asians (in yellow), and Overlap between whites and Asians (in black). Dark grey tracts are part of the city that do not belong to any of the zones, while light grey tracts are not part of New York City.

average per capita income of each race inside each zone for the years of 1990, 2000, and 2010. The results are presented in Fig 23, where "only" means the HD zone without the Overlap zone. The Figure shows that white citizens earn more than all the other races in all the zones except in the study of the segregation between white and Asian citizens. While, black and Hispanic citizens earn less than whites in all the zones. Moreover, the Figure shows that income inequality between white and black citizens is greater in the Overlap zone than in the only white zone and the only black zone.

To deepen the per capita income inequalities for each study of segregation (white and black, white and Hispanic, and white and Asian), we calculated the Gini coefficient [120] inside each of them. The results are presented in Fig 24. The Gini coefficient varies from 0 to 1. When it is next to 0, there is not inequality, while when it is next to 1, inequality is maximum [120]. The Figure shows that inequality is greater in the Overlap zones in all cases in favor of whites.

Tabela 3: Overlap Coefficients

	1990	**2000**	**2010**
White and Black	0.22	0.19	0.20
White and Hispanic	0.61	0.53	0.47
White and Asian	0.82	0.73	0.67
Black and Hispanic	0.52	0.52	0.61
Black and Asian	0.27	0.24	0.26
Hispanic and Asian	0.58	0.48	0.29

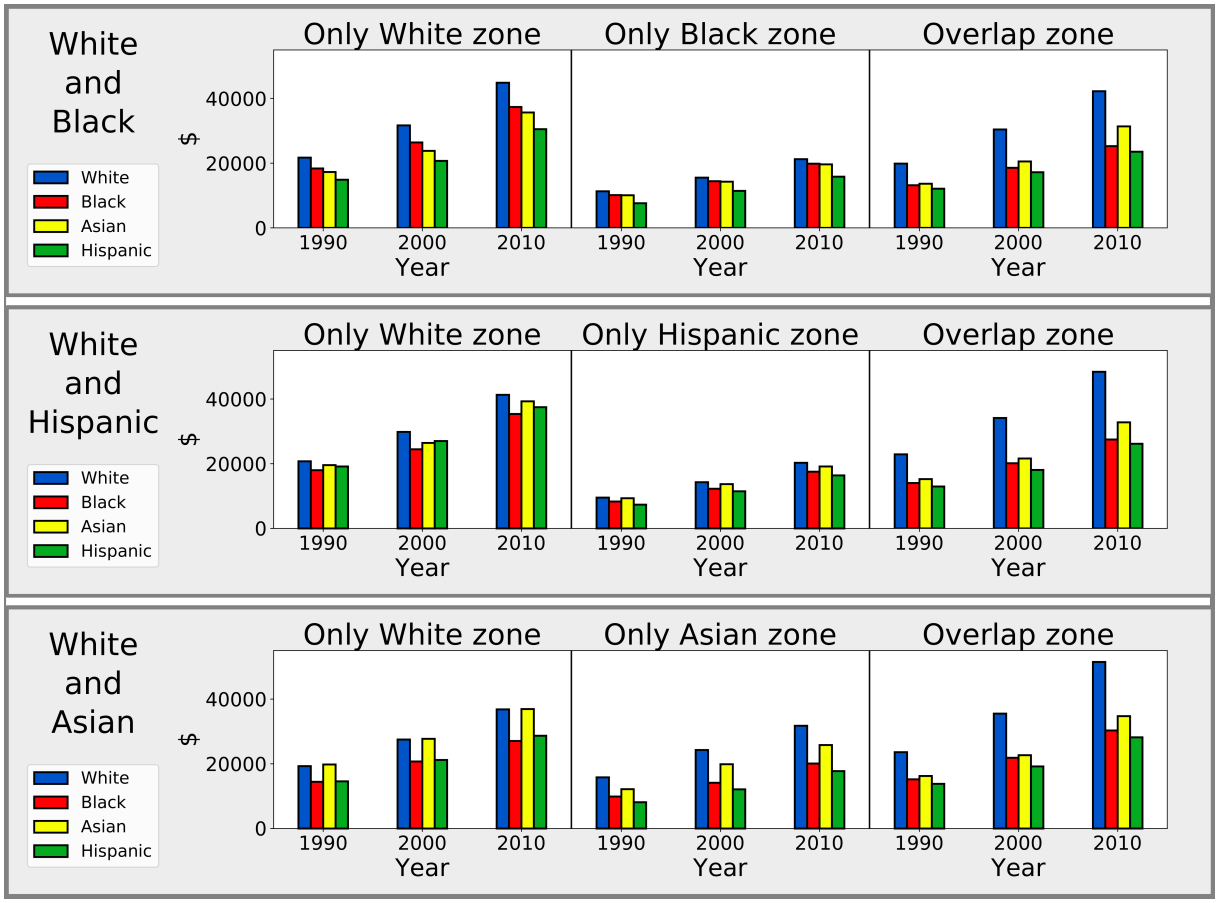

Figura 23: **Per capita income analysis.** The Figure shows the mean per capita income for each race for the study of the segregation between white and black, white and Hispanic, and white and Asian for the years of 1990, 2000, and 2010.

Furthermore, we analyzed the tracts that migrated from one zone to another from 1990 to 2010 for the studies of segregation between: white and black citizens in Fig 25; white and Asian citizens in Fig 26; and white and Hispanic citizens in Fig 27. The colors in the maps in Figs 25-26-27 show the alternatives of migration of the tracts from one zone to another, which are described in the caption. For each alternative, we calculated the average variation of the per capita income (ΔI) and the average variation of the properties values (ΔH) normalized by the average variation in the city ($\overline{\delta I}$ and

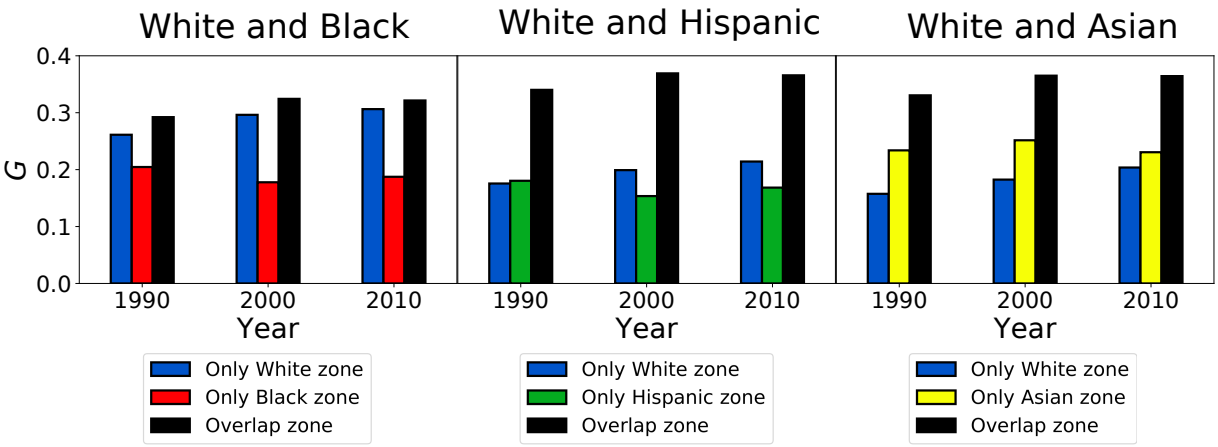

Figura 24: **Gini coefficient for the years of 1990, 2000, and 2010.** The Figure shows the Gini coefficient in the HD only zones and in the Overlap zones for the study of segregation between: white and black, white and Hispanic, and white and Asian.

$\overline{\delta H}$) from 1990 to 2010. The variations are defined as:

$$\Delta I = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta I_i - \overline{\delta I}}{\left|\overline{\delta I}\right|}, \tag{5.3}$$

and,

$$\Delta H = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta H_i - \overline{\delta H}}{\left|\overline{\delta H}\right|}. \tag{5.4}$$

Where N is the number of tracts of the analyzed pairs of races and δI_i and δH_i are the variations of the per capita income and properties values of tract i, respectively. Therefore, positive ΔI or ΔH mean growth higher than the city mean, while, conversely negative ΔI or ΔH mean growth lower than the city mean.

Moreover, we focused on the segregation between white and black citizens and the flux of people from 1990 to 2010 inside the tracts that migrated from one zone to another or to the Overlap zone. The flux of people of a specific race inside a tract is the variation of people of that specific race X inside tract i compared with the mean variation of that specific race in the whole city. Similarly to Eq 5.3 and 5.4, the average flux $\Delta Flux_X$ is defined:

$$\Delta Flux_X = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta Flux_{X,i} - \overline{\delta Flux_X}}{\left|\overline{\delta Flux_X}\right|}, \tag{5.5}$$

where $\overline{\delta Flux_X}$ is the mean flux of race X in the whole city.

In Fig 28, still focusing on the segregation between white and black citizens, we show: the variation of income; the variation of properties values; and the flux of people in the tracts that change zone between the years 1990 and 2010. For those tracts, in Fig 29 we compare the variation of the flux of white and black citizens with the variation of

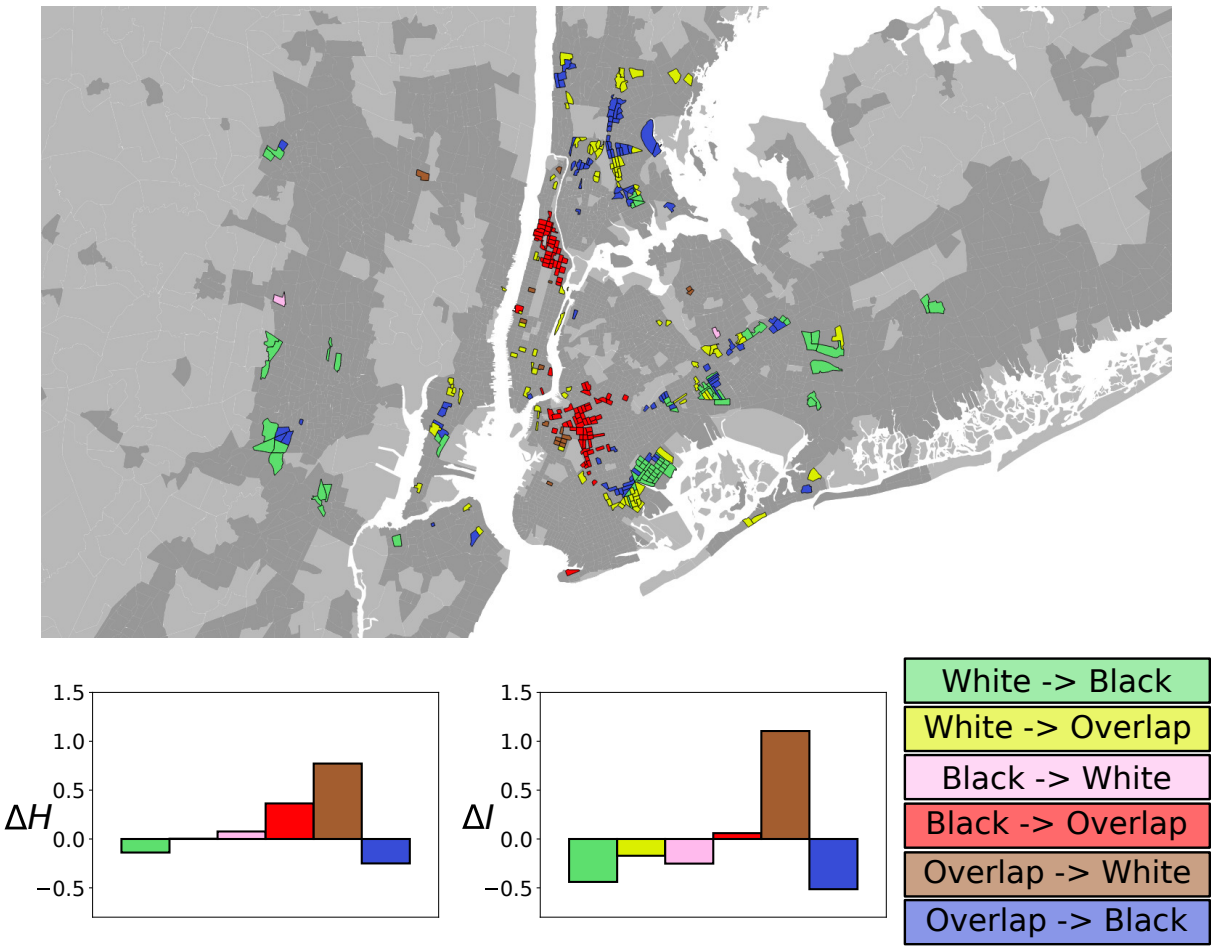

Figura 25: **Tracts that migrated from one zone to another or to the Overlap zone from 1990 to 2010: white and black citizens.** All the tracts that changed zone during the period from 1990 to 2010 are shown on the map, while the colors show the different alternatives of migration. Furthermore, for each alternative of migration, the value of ΔH and ΔI is shown.

the properties values. In Fig 29a, we show the outgoing white flux in orange where the red square is the centroid. In blue, we show the incoming white flux, where the black circle is the centroid. While in Fig 29b we show the outgoing black flux in green and the red square is the centroid. The incoming black flux in the considered tracts is shown in red and the black circle is the centroid. The Figures show that where the flux of white citizens is on average positive, also the properties values increase more than the mean, as well as where the flux of black citizens is negative on average.

To investigate the dynamics and the displacement of black citizens in New York City, we deepened the HD black zone. With a simplified version of the CCA we divided in clusters the HD black zone. Indeed, we ignored the threshold D^* and we apply the cutoff length ℓ'. The parameter ℓ' is chosen by analyzing the distribution of the tracts area. Each tract area is considered as a circle with the same area. The mean radius has been found to be $\bar{r} = 1.3\ km$, therefore in order to consider two neighbors tracts as part

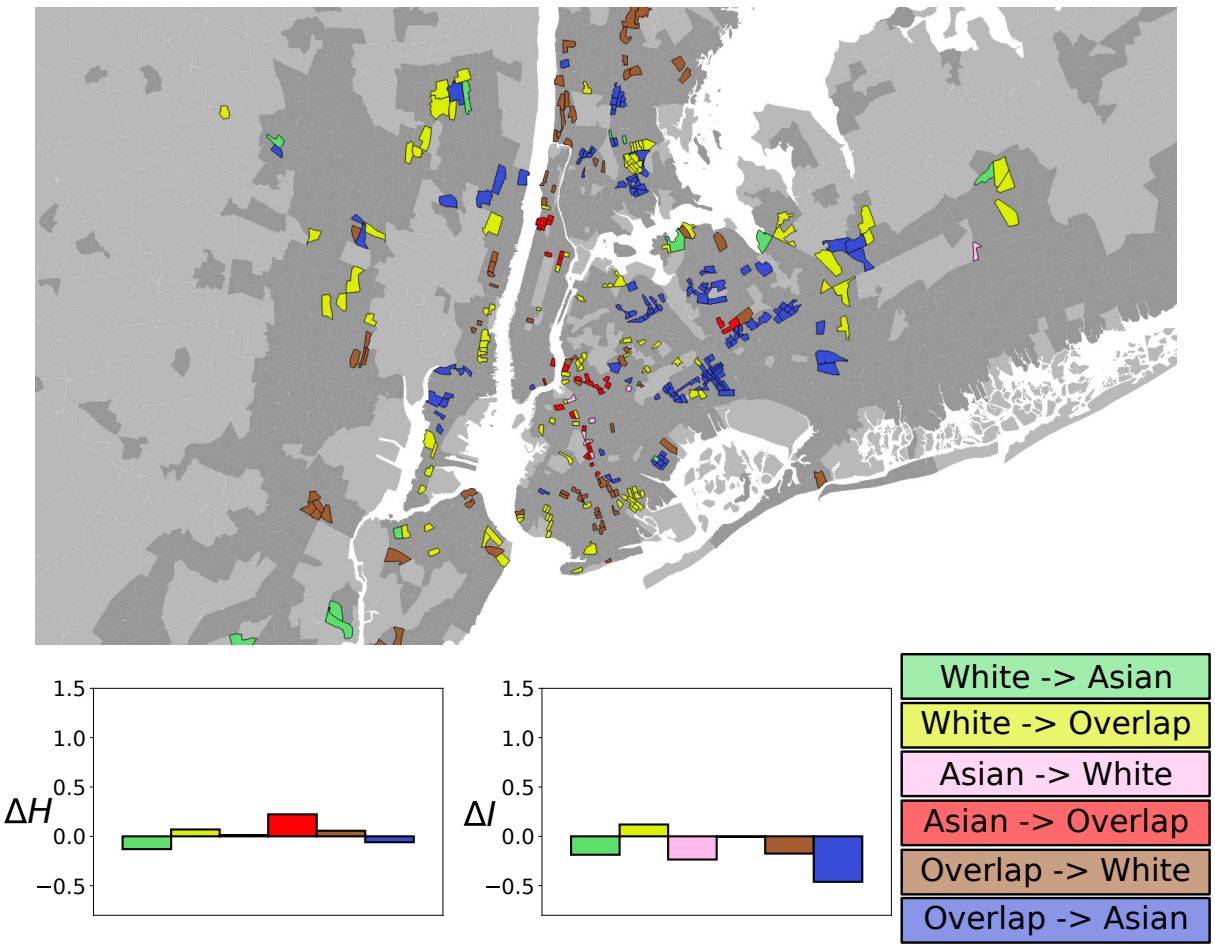

Figura 26: **Tracts that migrated from one zone to another or to the Overlap zone from 1990 to 2010: white and Asian citizens.** Similar to Fig 25, here we analyze white and Asian citizens.

of the same cluster, we used $\ell' = 1.5\ km$. The results of the clusterization for the years 1990 and 2010 are shown in Fig 30. In the Figure, we highlight the four biggest clusters A, B, C, and D.

For the four biggest clusters (A, B, C, and D), in Table 4 we show the area of each of them for the years 1990 and 2010 and also the displacement of clusters's centroid, highlighting the fact that cluster A and C have a displacement about three times higher than clusters B and D. In Fig 31, we show the displacement of clusters A and C from 1990 to 2010. The cluster A includes a region in the neighborhood of Harlem, while the cluster B is inside the boroughs of Brooklyn. In the same Figure, we also show the variation of the per capita income ΔI for the tracts that change zone in the analyzed period.

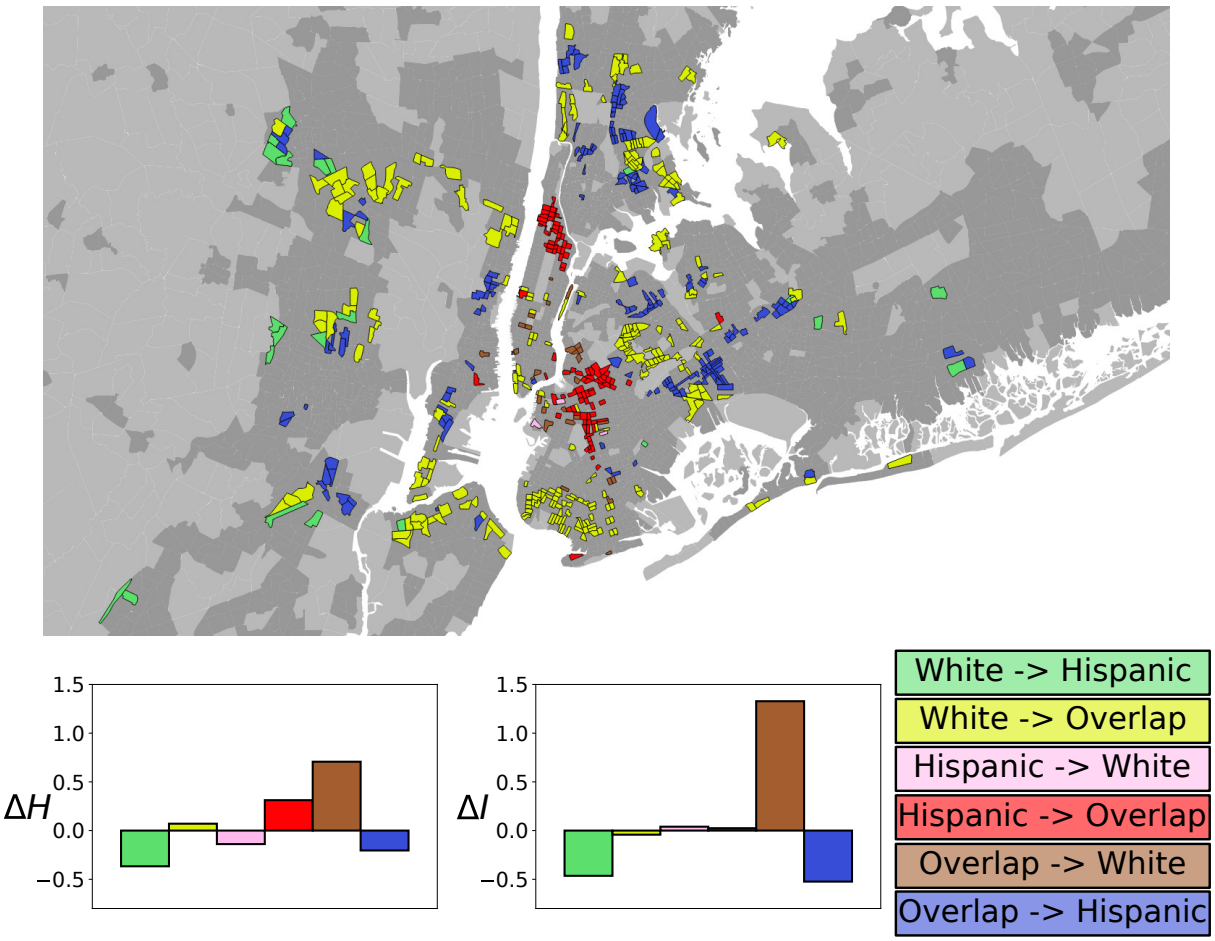

Figura 27: **Tracts that migrated from one zone to another or to the Overlap zone from 1990 to 2010: white and Hispanic citizens.** Similar to Fig 25 and 26, here we analyze white and Hispanic citizens.

5.5 Comparison with the Dissimilarity index

In order to verify the robustness of our method, we compared the Overlap coefficient defined in Eq 5.2 with the dissimilarity index [103]:

$$D_{ab} = \frac{1}{2}\sum_{i=1}^{N}\left|\frac{a_i}{A} - \frac{b_i}{B}\right|, \tag{5.6}$$

where a_i is the population of race a in tract i and b_i, the population of race b in the same tract. A and B are the total population of race a and b in the whole city, where the city

Tabela 4: Areas and displacements of the four biggest clusters of the HD black zone.

	$\mathbf{Area_{1990}}$ (km^2)	$\mathbf{Area_{2010}}$ (km^2)	$\mathbf{Displacement_{2010-1990}}$ (km)
A	30.7	32.8	1.55
B	38.0	54.6	0.44
C	41.8	44.1	1.57
D	37.3	58.2	0.64

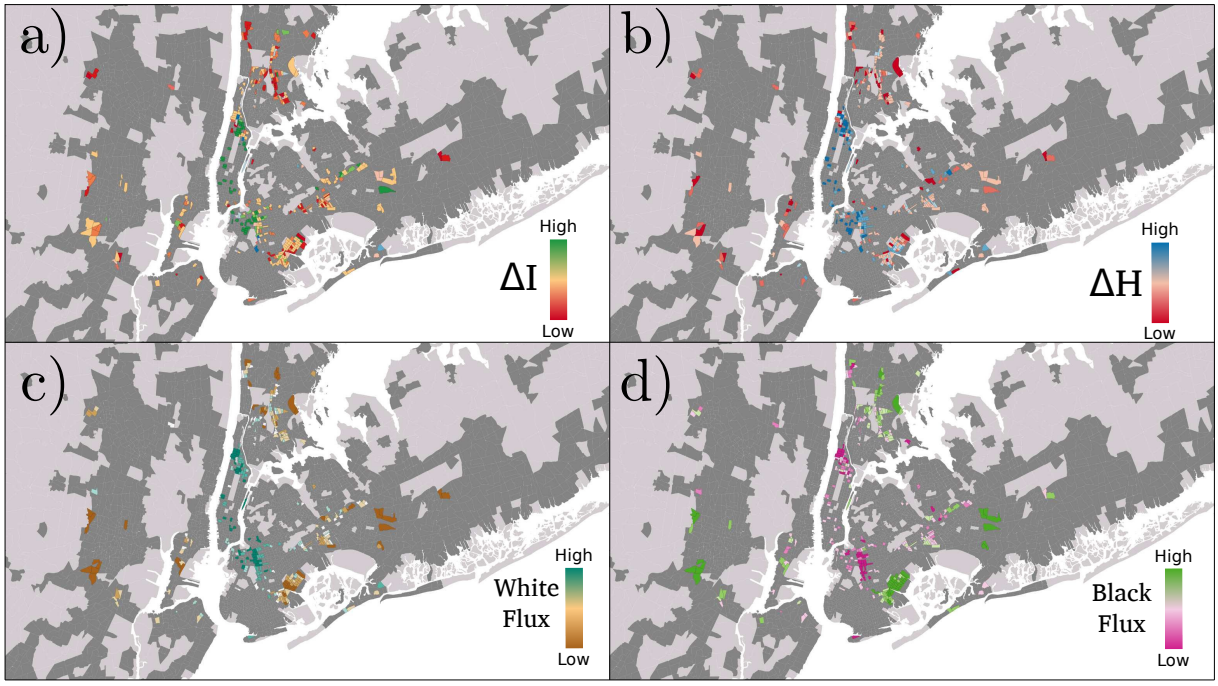

Figura 28: **Segregation between white and black.** The Figure shows: *a*) the variation of the per capita income, *b*) the variation of the properties values, *c*) the incoming flux of white, and *d*) the incoming flux of black for the tracts that migrated from one zone to another or to the Overlap zone from 1990 to 2010.

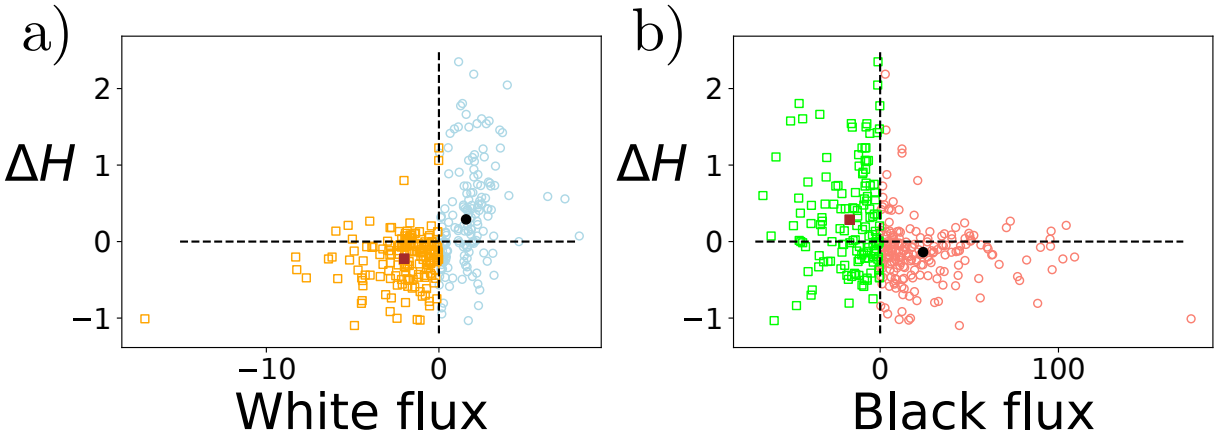

Figura 29: **Variation of properties values in function of the incoming flux of white and black citizens for the tracts that change zone from 1990 to 2010.** a) Variation of the properties values in function of the incoming flux of white citizens. The tracts with an outgoing flux of white are shown in orange, while the tracts with an incoming flux of white are shown in blue. The black red square is the centroid of the outgoing flux, while the black circle is the centroid of the incoming flux. b) Variation of the properties values in function of the incoming flux of black citizens. The tracts with an outgoing flux of black are shown in green, while the tracts with an incoming flux of black are shown in red. The black red square is the centroid of the outgoing flux, while the black circle is the centroid of the incoming flux.

is defined using the CCA. N are all the tracts that belong to New York City. The value of D_{ab} varies from 0 to 1. When it is next to 1, RRS is high, and vice versa, when it is

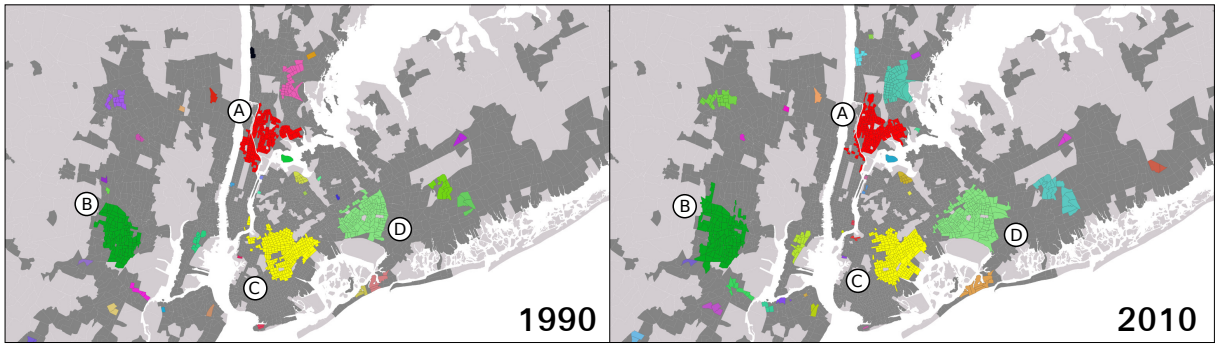

Figura 30: **Clusterization of the HD black zone for the years of 1990 and 2010.** The Figure shows the results of the clusterization of HD black zone using parameter $\ell' = 1.5\ km$ for the years of 1990 and 2010. The four biggest clusters A (in red), B (in dark green), C (in yellow), and D (in light green) are highlighted.

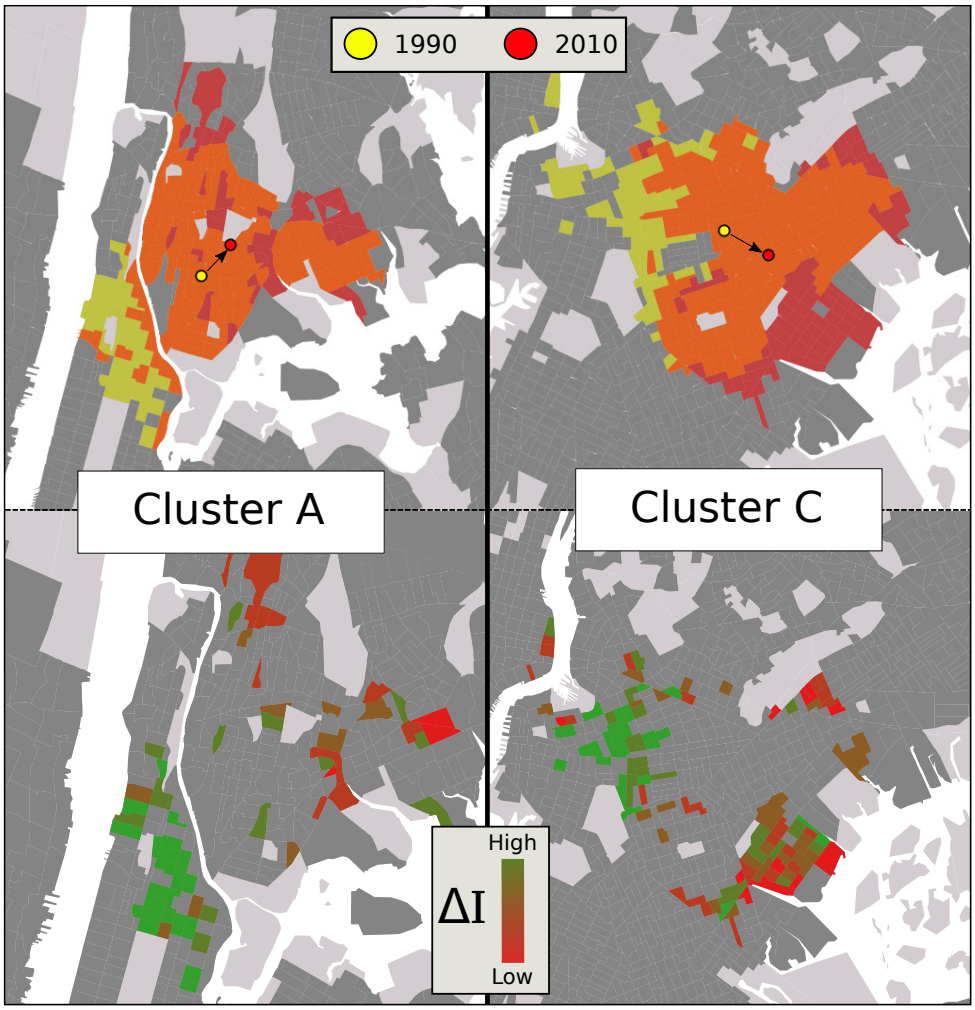

Figura 31: **Displacement of clusters A and C and the variation of per capita income.** The Figure shows the displacement of cluster A (equivalent to the neighborhood of Harlem and the borough of Bronx) and C (equivalent to the borough of Brooklyn). The clusters in the year 1990 are shown in yellow and the clusters in the year of 2010 are shown in red, with the respective centroids. The figures below show qualitatively the variation of the per capita income for the tracts that change zone in the analyzed period.

Tabela 5: Dissimilarity index

	1990	**2000**	**2010**
White and Black	0.81	0.80	0.79
White and Hispanic	0.64	0.64	0.62
White and Asian	0.47	0.50	0.51
Black and Hispanic	0.58	0.58	0.54
Black and Asian	0.78	0.78	0.76
Hispanic and Asian	0.56	0.58	0.58

next to 0 there is not segregation. It shows the percentage of one of the two populations that have to move in order to reduce segregation to 0 [103]. The results obtained in New York City are shown in Table 5.

To analyze the correlation between the two indexes, we plot the dissimilarity indexes D_{ab} found in New York City as a function of their respective Overlap coefficients $O_{rr'}$ (where X_r is the HD zone of race a, and $X_{r'}$ of race b) in Fig 32. The red line in the Figure shows the result of the Ordinary least Square (OLS). As expected, the relation is inverse with a linear coefficient $m = -0.57 \pm 0.01$. Whereupon, in order to quantify the correlation between the two indexes, we calculated the Pearson correlation coefficient (PCC), $\rho_{D,O} = -0.96$. The value implies a strong inverse correlation between the two indexes, proving the robustness of our method.

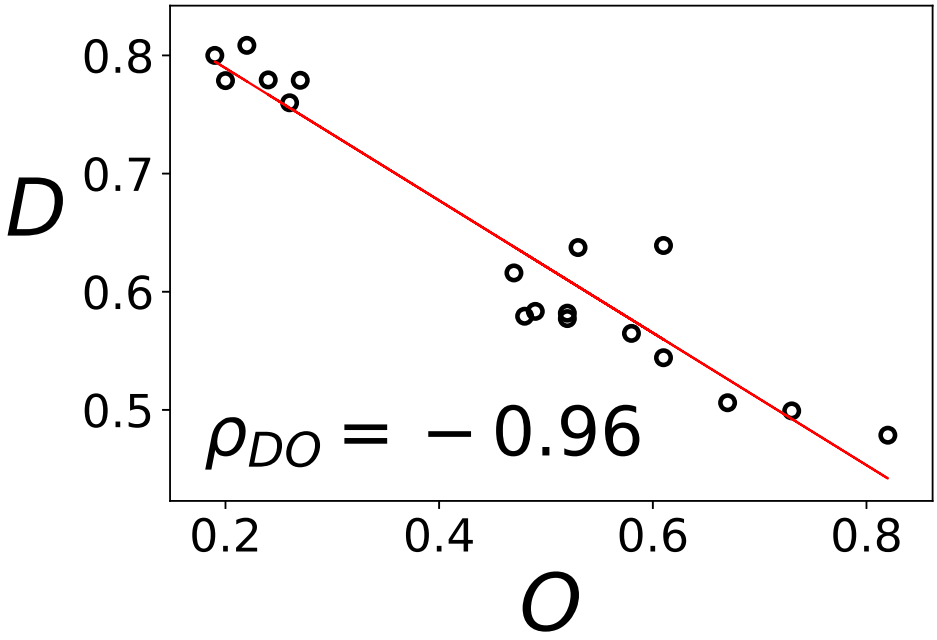

Figura 32: **Dissimilarity index D as a function of the Overlap coefficient O.** The red line is the OLS with angular coefficient $m = -0.57 \pm 0.1$. The Pearson correlation coefficient, $\rho = -0.96$, shows a strong inverse correlation between the two indexes.

5.6 Discussion

We developed a new method in order to measure and to define the topography of RRS and it has been applied to the metropolitan area of New York City for the years of 1990, 2000, and 2010. Despite the fact that several studies show that, on average, segregation between white and black citizens in the United States has decreased in the last fifty years [91–94], our results show that it has remained quite stable during the time interval 1990-2010 in the metropolitan area of New York City as well as for black and Asian citizens. Instead, segregation between white and Hispanic, white and Asian, and Hispanic and Asian citizens has grown. Only black and Hispanic are less segregated in 2010 compared with 1990.

By analyzing the per capita income, we observed that white citizens earn more than the other races in all the regions, except when we studied the segregation between whites and Asian, where Asian citizens have a similar income to white citizens. Regarding the segregation between white and black citizens, we verified that black citizens earn less than white citizens in all the regions. Furthermore, the inequality between white and black citizens is greater in the regions of high density of population of both the races. This result is confirmed by the Gini coefficient, in fact we showed that it is higher in the regions of high density of population of two or more races.

Furthermore, we deepened the segregation between white and black and the segregation between white and Hispanic citizens. We analyzed the tracts that change population density from 1990 to 2010 (from region of high density of black, Hispanic, or overlap with white citizens) to region of only high density of white citizens. In this region, we observed that the per capita income and the properties values increased more than the city mean. Conversely, in the tracts that migrated from a region of overlap to a region with high density of population of only black or Hispanic citizens we observed that the per capita income and the properties values increased less than the mean. The same does not happen deepening the segregation between white and Asian citizens.

Focusing on the segregation between white and black citizens, we analyzed the flux of white and black citizens in function of the variation of the properties values. Here, we confirmed our previous result, that is where the flux of white citizens is positive, the properties values increased more than the city mean, while, where the flux of black citizens is positive, the properties values increased less than the city mean. How can low-income black citizens continue to live in places where the properties increase more than the city mean?

Previous studies [116–118] questioned the effects of gentrification in the neigh-

borhood of Harlem and in the borough of Brooklyn. Here, by clustering the region of high density of black citizens, we showed the displacement of the clusters defined as A (that include a region inside the neighborhood of Harlem) and B (that is inside the borough of Brooklyn). The displacement is of respectively 1.55 km and 1.57 km in twenty years. This result confirms the theory of displacement of black citizens in the neighborhood of Harlem and in the borough of Brooklyn. Moreover, we showed the census tracts that migrate out from the cluster A or C have an increase of the per capita income higher than for the tracts that migrate in those clusters.

6 GENERAL CONCLUSION

The aim of this thesis is to show three projects developed by us in the field of the physics of complex systems. All the three projects are founded in the statistical analysis of real data. The first project *The light pollution as a surrogate for urban population of the US cities* showed that it is plausible to utilize night-time light as a surrogate for city population. Furthermore, it showed that there is no economy of scale or sublinearity concerning the night-time light in US cities and we corroborated previous works showing that the scaling behaviors of urban indicators with population can be substantially different for distinct definitions of city boundaries.

In the second project *Dynamics in the Fitness-Income plane: Brazilian States VS World Countries*, we developed a variant of the Fitness algorithm for subnational entities and we applied it to the Brazilian states. Our results showed deep analogies between the dynamics of the world countries and the dynamics of the Brazilian states in the Fitness-Income plane. Indeed, we showed the high predictability of growth of the economy of several states, while for others the dynamics is less predictable.

In the last project entitled *Dynamics of racial segregation and gentrification in New York City*, we analyzed the phenomenon of the racial residential segregation in New York City. We developed a new index for measure the residential segregation able to define the topography of the segregated zones. We compared the dynamics of the segregation with the dynamics of per capita income and the properties values, showing deep discrepancies in function of the region and the races. Furthermore, we measured the displacement of black citizens in the gentrified neighborhood of Harlem and in the borough of Brooklyn.

These three projects are examples of how physicists can contribute in social sciences and in economy. The ability of extract information from a large and, often complex, data is not only a peculiarity of physicists, but it needs the collaboration and the interconnection among scientists of different areas. This view is coherent with an holistic view of the nature, where divisions inside the science are only human artifices. Perhaps, the state of the art of a new science is without limits among the fields of study, and in this context the physics of the complex systems is a little step in this direction.

APÊNDICE A – PUBLISHED PAPERS

ARTICLE IN PRESS

Physica A ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

The light pollution as a surrogate for urban population of the US cities

Felipe G. Operti [a], Erneson A. Oliveira [a,b,*], Humberto A. Carmona [a], Javam C. Machado [c], José S. Andrade Jr. [a]

[a] Departamento de Física, Campus do Pici, Universidade Federal do Ceará, 60451-970, Fortaleza, Ceará, Brazil
[b] Programa de Pós-Graduação em Informática Aplicada, Universidade de Fortaleza, 60811-905, Fortaleza, Ceará, Brazil
[c] Departamento de Computação, Campus do Pici, Universidade Federal do Ceará, 60455-760, Fortaleza, Ceará, Brazil

HIGHLIGHTS

- City boundaries influence the scaling of the night light as a function of population.
- Small and large cities are indistinguishable in terms of light pollution.
- It is plausible to utilize the night-time light as a surrogate for city population.

ARTICLE INFO

Article history:
Received 14 June 2017
Received in revised form 1 November 2017
Available online xxxx

MSC:
00-01
99-00

Keywords:
Allometry
Night-time light
Light pollution
City clustering algorithm
Metropolitan/Consolidated Metropolitan Statistical Area

ABSTRACT

We show that the definition of the city boundaries can have a dramatic influence on the scaling behavior of the night-time light (NTL) as a function of population (POP) in the US. Precisely, our results show that the arbitrary geopolitical definition based on the Metropolitan/Consolidated Metropolitan Statistical Areas (MSA/CMSA) leads to a sublinear power-law growth of NTL with POP. On the other hand, when cities are defined according to a more natural agglomeration criteria, namely, the City Clustering Algorithm (CCA), an isometric relation emerges between NTL and population. This discrepancy is compatible with results from previous works showing that the scaling behaviors of various urban indicators with population can be substantially different for distinct definitions of city boundaries. Moreover, considering the CCA definition as more adequate than the MSA/CMSA one because the former does not violate the expected extensivity between land population and area of their generated clusters, we conclude that, without loss of generality, the CCA measures of light pollution and population could be interchangeably utilized in future studies.

1. Introduction

More than 80% of the world and more than 90% of the US and European populations live under light-polluted skies (exposition to light at night) [1]. Since the first electric-powered illumination in the second half of the 19th century, the world has become covered by artificial electric light, changing drastically the night view of the Earth from space. The spreading of artificial electric light plays an important role on the duration of the *productive day*, not only for working but also for recreational activities. If in one hand the benefits of artificial light are quite evident, on the other hand, scientific researches

* Corresponding author at: Departamento de Física, Campus do Pici, Universidade Federal do Ceará, 60451-970, Fortaleza, Ceará, Brazil.
E-mail address: erneson@fisica.ufc.br (E.A. Oliveira).

https://doi.org/10.1016/j.physa.2017.11.039
0378-4371/

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

suggest that the exposition to light at night could have adverse effects on both human and wildlife health [2–9]. For example, in humans, the pineal and blood melatonin rhythms are quickly disturbed by light pollution. Such studies argue that the night light exposure have two major physiological effects: they disrupt the circadian rhythms and suppress the production of melatonin [8]. This repeated suppression may have large consequences for the mammals health. For instance, it was shown that the suppression of the melatonin at night accelerates the metabolic activity and growth of rat hepatoma [5] and human breast cancer [3]. Moreover, the disruption of circadian rhythms made by the exposure of light at night might plays a crucial role in the etiology of depression [8].

The significant consequences of the exposure to night-time light (NTL) with the fact that 54% of world's population lives in urban areas stimulates the interest in understanding how the light pollution evolves with the size of the US cities [10]. Bettencourt et al. found the cities in the US exhibit three different types of allometric laws for urban indicators with population size [11]: (i) *Superlinear* . The superlinear urban indicators increase proportionally more than the population of the cities. Such behavior is intrinsically associated with the *social currency* of a city, indicating that larger cities are associated with optimal levels of human productivity and quality of life. Doubling the city size leads to a larger-than-double increment in productivity and life standards [11–13]. For example, wages, income, growth domestic product (GDP), bank deposits, as well as rates of invention measured by the number of patents and employment in creative sectors show a superlinear behavior [11]. (ii) *Linear* or *isometric relation*. The increasing of the linear urban indicators is proportional to the increasing of the population reflecting the common individual human needs, like the number of jobs, houses, and water consumption [11]. (iii) *Sublinear*. The sublinear urban indicators increase proportionally less than the population of the cities. This case is a manifestation of the *economy of scale*. The sublinearity is found in the number of gasoline stations, length of electrical cables, and road surfaces (material and infrastructure) cases [11]. From the results shown by Bettencourt et al., several studies have been carried out on the allometry of urban indicators in different levels of human aggregation [14–20]. For instance, recently there were found correlations between the flows of energies and material (such as electricity consumption, water consumption, etc.) and several urban indicators (such as population growth, economic activity, etc.) for the world's 27 megacities with populations greater than 10 million people in 2010 [21]. Furthermore, the deepening of the energy metabolism of megacities showed an allometric scaling between the per capita total energy consumption and the urban population density with the characterized -3/4 coefficient, proving that compact cities are more energy efficient with respect dispersed ones [22]. Following this aim, we analyze and classify the allometric law between the NTL and the population of the US cities.

Here, we use three geo-referenced dataset: the population dataset, the NTL dataset and the Metropolitan/Consolidated Metropolitan Statistical Area (MSA/CMSA). In order to define the boundaries of each US city, we use two methods: the City Clustering Algorithm (CCA) [23–26] and the MSA/CMSA [27]. Finally, we find the allometric scaling between the NTL and the population for the two applied methods. Furthermore, to compare them, we analyze the allometric scaling between area and population.

2. Materials and method

2.1. Population dataset (GPWv4)

The population dataset is extracted from the fourth version of the Gridded Population of the World (GPWv4) [28,29] from the Center for International Earth Science Information Network (CIESIN) at the Columbia University. The GPWv4 models the human population distribution on a continuous surface at high resolution. Population input data is collected through several censuses around the US, between 2005 and 2014. Data are provided in grid form, where each cell is formed by 30 arc-second angles (approximately 1 km $\times$ 1 km at the Equator line). We use the US population count data, measured in number of people, for the year 2015, as depicted in Fig. 1a.

The method successively introduced requires the population density of each grid cell. Therefore, we calculated the area of each grid cell dividing them into two spherical triangles. The area of a spherical triangle with edges a, b and c is given by,

$$A = R^2 E, \tag{1}$$

where $R = 6378.137$ km is the Earth's radius and the spherical excess E is defined by the following expression:

$$E = 4\tan^{-1}\left[\tan\left(\frac{s}{2}\right)\tan\left(\frac{s_a}{2}\right)\tan\left(\frac{s_b}{2}\right)\tan\left(\frac{s_c}{2}\right)\right]^{1/2}. \tag{2}$$

with $s = (a/R + b/R + c/R)/2$, $s_a = s - a/R$, $s_b = s - b/R$, and $s_c = s - c/R$. In this context, the distance between two points, i and j, on the Earth's surface is calculated by,

$$d_{ij} = R\theta, \tag{3}$$

with

$$\theta = \cos^{-1}[\sin(y_i)\sin(y_j) + \cos(y_i)\cos(y_j)\cos(x_j - x_i)]. \tag{4}$$

In this formalism, the values of x_i (x_j) and y_i (y_j) are the longitude and latitude, respectively, of the point i (j), measured in radians.

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

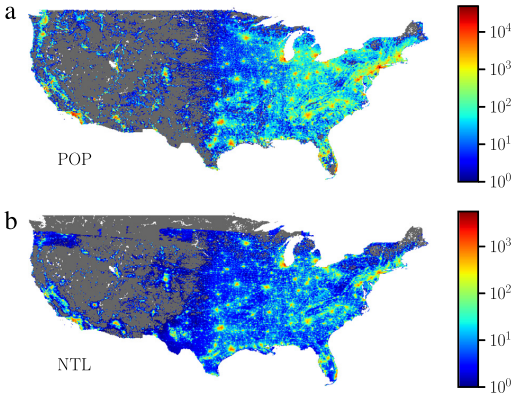

Fig. 1. Datasets (on colors). (a) The population dataset is defined as a 30 arc-second geolocated grid. It is obtained from the GPWv4 in logarithmic scale for the year 2015 [28,29]. (b) The NTL dataset is obtained through the night-time light radiance emission data from the VIIRS DNB in nW/cm^2/sr [30–32]. It is defined at the resolution of 15 arc-second grid in logarithmic scale for the year 2015 (April).

2.2. Night-time light dataset (NTL)

The NTL dataset is given by the night-time light radiance emission data from the National Centers for Environmental Information (NCEI) [31]. The NTL dataset is defined by the monthly average of radiance, measured in nW/cm^2/sr, using the night-time data from the scanning radiometer Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) [30–32]. The VIIRS DNB data are processed and filtered in order to exclude data impacted by the lunar illumination, lightning and cloud-cover, but they are susceptible to other temporal lights, *e.g.* aurora, fires, and boats [30,31]. Such data span through the entire globe with a resolution of 15 arc-second (approximately 500 m $\times$ 500 m at the Equator line) between the latitudes 75° North and 65° South. We use the US data for the year 2015 (April), as shown in Fig. 1b.

2.3. Metropolitan Statistical Area (MSA), Primary Metropolitan Statistical Area (PMSA) and Consolidated Metropolitan Statistical Area (CMSA)

The MSA are geographic entities with high degree of socioeconomic integration and population over 50,000 people. The PMSA are quite similar to MSA, however they present population over 1,000,000 people. The CMSA are metropolitan regions defined by the agglomeration of some PMSA. They are all delineated by the Office of Management and Budget (OMB) and provided by the US Census Bureau [27].

2.4. Data processing

In order to superimpose the datasets, we perform two processes: (i) The two datasets, NTL and GPWv4, have different resolutions. Indeed, the NTL grid has a higher resolution than the GPWv4 grid. Here, we define a new NTL grid with the same positioning and resolution of the GPWv4 dataset. Therefore, the new NTL cells are defined by the adding of the inner old NTL cells, *i.e.* the old NTL cells within the perimeter of each new NTL cell. (ii) For the MSA/CMSA case, we use the same approach of (i), even though the MSA/CMSA are complex polygons. To deal with this problem, we use the even–odd rule algorithm [33]. Thus, we define the NTL value for each MSA/CMSA.

2.5. City Clustering Algorithm (CCA).

We define the boundaries of each US city by applying the CCA to the population grid [23–26]. We use the continuum CCA that depends on two parameters, namely, a population density threshold, D^* and a cutoff length, ℓ [26]. For the ith grid cell, the population density D_i is geo-referenced in its geometric center (shown as small black circles in Fig. 2). If $D_i > D^*$, the ith grid cell is populated. In Fig. 2 the populated cells are shown in gray and red. Next, the algorithm selects a populated cell (red cell in Fig. 2a) and aggregates in the same cluster all nearest populated cells which are within a distance ℓ from each (red cells in Figs. 2b, c and d). The Fig. 2 shows the four steps to determine the red cluster.

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

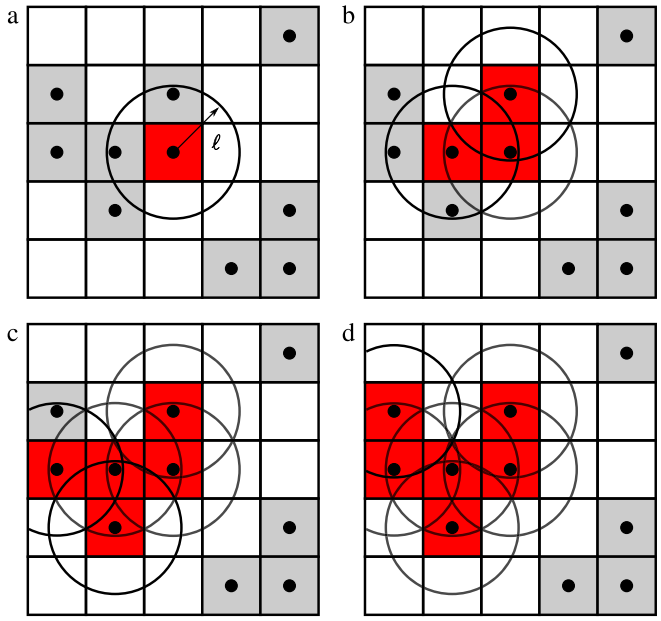

Fig. 2. The CCA steps. The gray and the red cells are populated ($D_i > D^*$). The small black circles are the geometric centers of each populated cell. The red cells belong the same analyzed cluster. (a) First step: the algorithm select a populated cell and draw a circle of radius ℓ. (b) Second step: the cells with the geometric centers inside the circles of radius ℓ become a part of the red cluster and from their geometric center are drawn others two circles of radius ℓ. The circle of the first step is showed in opaque black. (c) Third step: two more cells became part of the red cluster and two more circles are drawn. (d) Fourth step: the last cell became part of the red cluster. The entire cluster is determined and the algorithm will start to analyze another cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Results

We apply the CCA to the population grid varying D^* (in people/km^2), from 0 to 10 000, and ℓ (in km), from 1 to 20. For all pairs of parameters, we find that it is possible to statistically correlate through power-law relations the area and the population as well as the NTL and the population of the US cities,

$$\log(\text{AREA}) = a + \alpha_{CCA} \log(\text{POP}), \tag{5}$$

$$\log(\text{NTL}) = b + \beta_{CCA} \log(\text{POP}). \tag{6}$$

The exponents α_{CCA} and β_{CCA} are obtained through Ordinary Least Square (OLS) [34] fitting to the data for different values of the parameters D^* and ℓ. The ranges of compatibility and the consistency of the CCA technique are investigated in Fig. 3a–d.

Indeed, the definition of the parameters D^* and ℓ of the CCA affects the dimension and the geometry of the cities, but from Figs. 3c and d, it can be seen that it does not affect the allometric exponent β_{CCA}. In Fig. 3d, there is a noticeable tendency towards smaller values of β for $\ell > 10$. This fact is due to the grouping of many clusters at once, which leads to a decreasing of the number of cluster samples. Therefore, such decreasing is reflected in large error bars for $\ell > 10$, *i.e.* a larger statistical fluctuation. Here, our starting strategy is to determine a range of parameters D^* and ℓ for which the relation between area and population is isometric [25,26,15,19]. We find that for $D^* > 4000$ and $\ell = 3$ the allometric exponent α_{CCA} is between 0.93 and 0.95 and we consider this relation approximately linear. Inside this range, we analyze the result of the CCA using $D^* = 4560$ and $\ell = 3$, where the five larger cities in the US Northeast Coast naturally emerge, as depicted in Fig. 4. We believe that, the lack of an exactly linearity, also inside this range, is due to the high density of some downtowns, specifically, of the most populated urban centers of the US Northeast Coast.

For the pair of parameters, $D^* = 4560$ and $\ell = 3$, we find a allometric exponent $\alpha_{CCA} = 0.93 \pm 0.01$ (Fig. 5a). The slight sublinearity of the scaling exponent is due to the lack of statistics for large cities and, consequently, such fact tends to produce a small deviation in the linear regression coefficient. We also find a linear scaling between NTL and the population

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

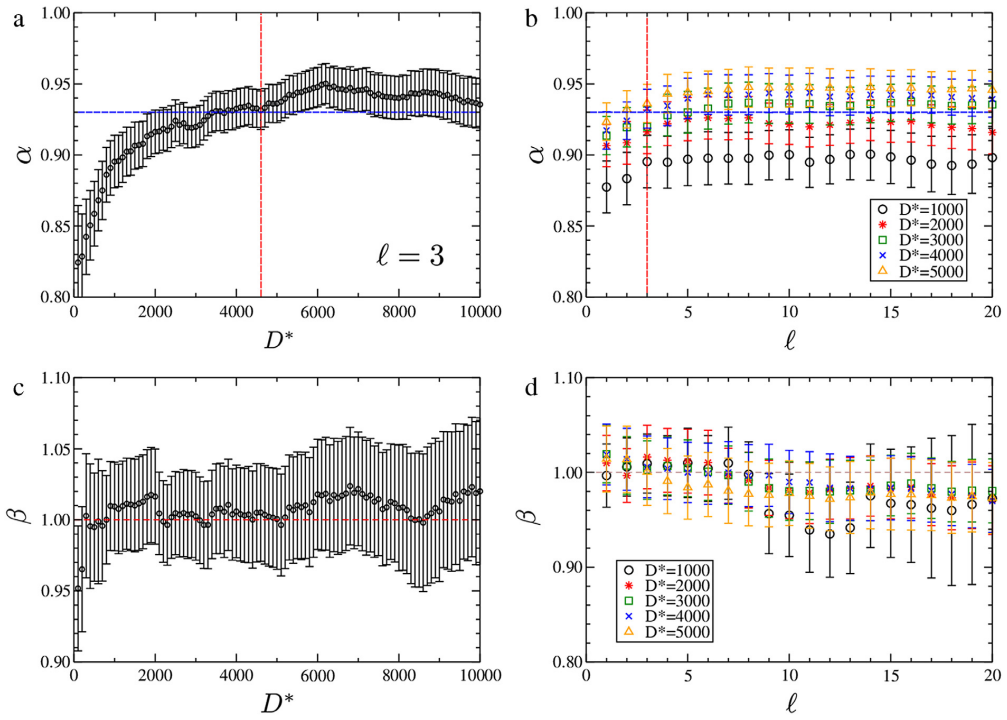

Fig. 3. Allometric exponent α_{CCA} and β_{CCA} as a function of the parameter D^* and ℓ. (a) The exponent α_{CCA} as a function of D^* for $\ell = 3$ km. The parameter D^* varies from 0 to 10000 people/km^2. For $D^* > 4000$ and $\ell = 3$ the allometric exponent α_{CCA} is between 0.93 (dashed blue line) and 0.95. For $D^* = 4560$ people/km^2 (dashed red line) we observe the arising of five large cities in US Northeast Coast. (b) The exponent α_{CCA} as a function of the CCA parameter ℓ for $D^* = 1000$, 2000, 3000, 4000, and 5000 people/km^2. We find a plateau region after $\ell = 3$ km, where $\alpha_{CCA} \approx 0.93$ (dashed blue line). (c) The figure shows the allometric exponent β_{CCA} as a function of D^* for $\ell = 3$ km. The parameter D^* varies from 0 to 10000 people/km^2. The dashed red line corresponds to $\beta = 1$. (d) The figure shows the allometric exponent β_{CCA} as a function of the CCA parameter ℓ for $D^* = 1000$, 2000, 3000, 4000, and 5000 people/km^2. The dashed brown line corresponds to $\beta = 1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

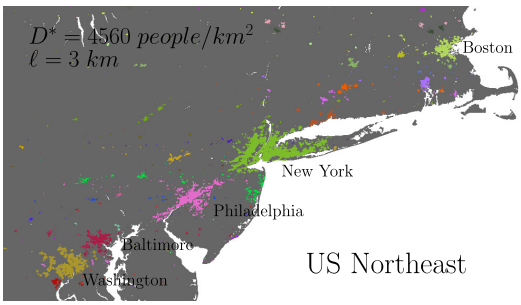

Fig. 4. Application of CCA to the US Northeast region. We use the CCA parameters $D^* = 4560$ people/km^2 and $\ell = 3$ km. The clusters of different colors identify different urban agglomerations. Essentially, we distinguish five famous cities such as Boston (light green), New York (green), Philadelphia (pink), Baltimore (red), and Washington D.C. (gold). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

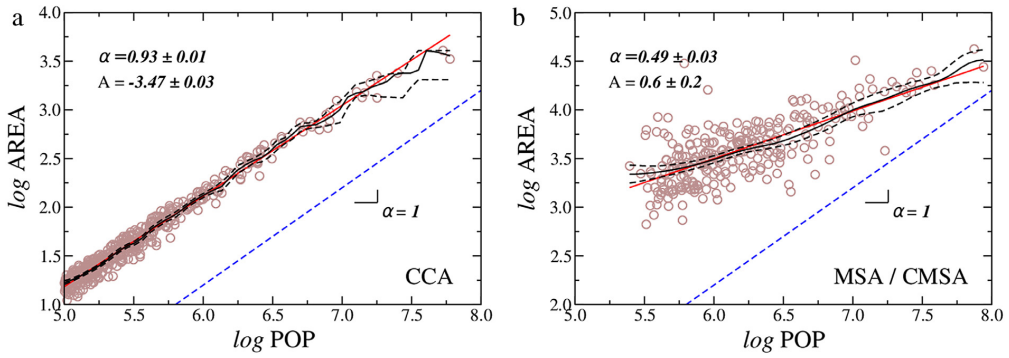

Fig. 5. Allometric exponent α applying the CCA and using the MSA/CMSA definitions. (a) The figure shows the allometric scaling law in Eq. (5) and its allometric scaling exponent $\alpha_{CCA} = 0.93 \pm 0.01$ using CCA parameters $D^* = 4560$ people/km^2 and $l = 3$ km. The red line is the OLS result, and the solid black line is the Nadaraya–Watson estimator (N–W) [35,36]. The dashed black lines show the 95% confidence bands of the N–W. The dashed blue line corresponds to $\alpha = 1$. (b) The figure shows the allometric scaling exponent $\alpha_{MSA/CMSA} = 0.49 \pm 0.03$ using the MSA/CMSA definitions. The red line is the OLS result, and the solid black line is the N–W estimator. The dashed black lines show the 95% confidence bands of the N–W. The dashed blue line corresponds to $\alpha = 1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

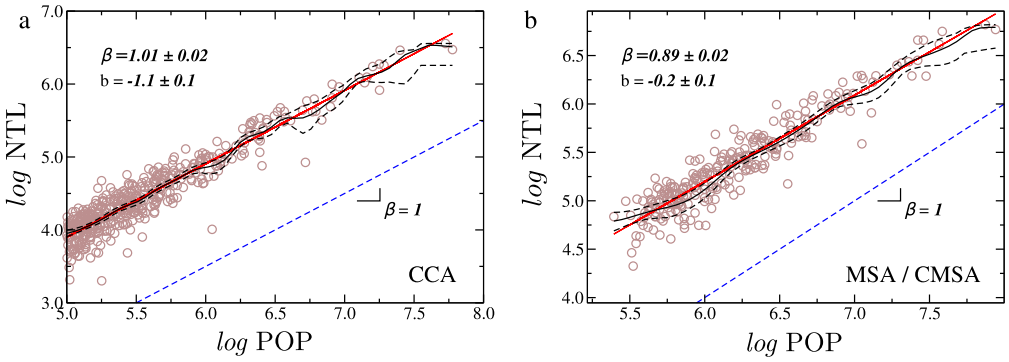

Fig. 6. NTL versus population using the CCA and the MSA/CMSA definitions. (a) NTL versus population using CCA parameters $D^* = 4560$ people/km^2 and $l = 3$ km. The graph shows a linear relation between the NTL measured in nW/cm^2/sr and the population with allometric scaling exponent $\beta_{CCA} = 1.01 \pm 0.02$ ($R^2 = 0.88$). The solid red line is the linear regression obtained using the OLS method. The solid black line is the Nadaraya–Watson estimator (N–W) and the dashed black lines show the lower and the upper confidence interval (95%) [35,36]. The dashed blue line corresponds to $\beta = 1$. (b) NTL versus population using MSA/CMSA. The graph shows a sublinear relation between the NTL, measured in nW/cm^2/sr, and the population with allometric scaling exponent $\beta = 0.89 \pm 0.02$ ($R^2 = 0.89$). The red line is the linear regression and the black line is the N–W estimator. The dashed black lines show the 95% confidence band of the N–W. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with exponent $\beta_{CCA} = 1.01 \pm 0.02$ (Fig. 6a). Alternatively, others parameters inside this range could be analyzed without affecting the allometric exponent β_{CCA} (as shown in Figs. 3c and 3d).

By analyzing the allometric scaling of the NTL with the population of the US cities using the MSA/CMSA (Fig. 6b), we obtain the allometric exponent $\beta_{MSA/CMSA} = 0.89 \pm 0.02$. Such an exponent characterizes a sublinear relation between the NTL and the population, in contrast with the CCA result.

As shown in Fig. 5b, the sublinear scaling behavior of the MSA/CMSA areas as a function of their corresponding populations, $\alpha_{MSA/CMSA} = 0.49 \pm 0.03$, clearly suggests that this might not be the most adequate definition of a city agglomerate to be adopted in our study.

As indicated by Oliveira et al. [15], the arbitrary geopolitical concept behind the MSA/CMSA seems to overestimate the natural limits of urban areas. In order to illustrate this fact, we show in Fig. 7 the MSA/CMSA of the five most populated US regions, namely, New York-Northern New Jersey-Long Island (NY, NJ, CT, PA), Los Angeles-Riverside-Orange County (CA), Chicago-Gary-Kenosha (IL, IN, WI) and Houston-Galveston-Brazoria (TX). The first and second columns show respectively

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

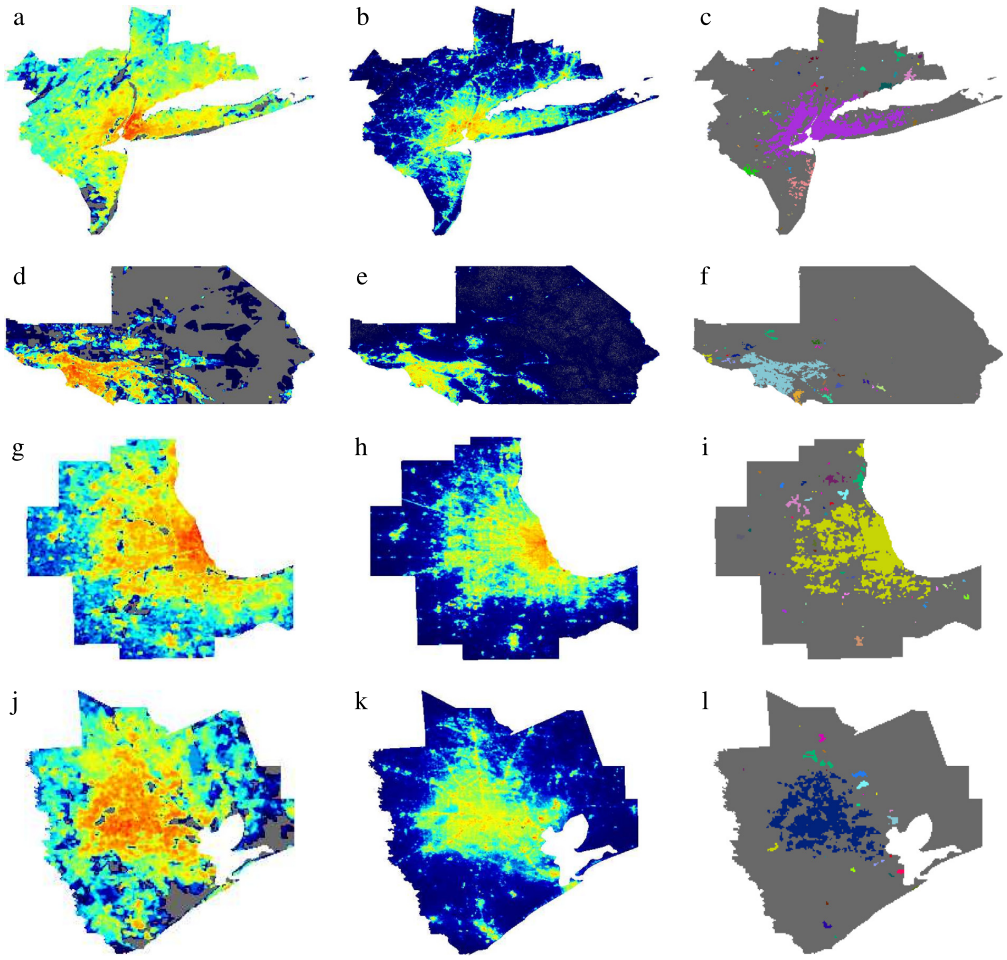

Fig. 7. Comparison between the CCA and MSA/CMSA (on colors). Figures (a), (d), (g) and (j) are the human population grid in logarithmic scale obtained from the GPWv4 for the year 2015 [28,29]. Figures (b), (e), (h) and (k) are the NTL measured in logarithmic scale with units nW/cm^2/sr obtained through the night-time light radiance emission data from the VIIRS DNB [30–32]. In figures (c), (f), (i) and (l) we show the CCA clusters obtained using the CCA parameters $D^* = 4560$ people/km^2 and $l = 3$ km of the CMSA of: New York-Northern New Jersey-Long Island (NY, NJ, CT, PA), Los Angeles-Riverside-Orange County (CA), Chicago-Gary-Kenosha (IL,IN,WI) and Houston-Galveston-Brazoria (TX). The figures show the discrepancy between the area estimated by the MSA/CMSA and the area delimited by the CCA.

the detailed maps of the population and the NTL datasets. The third column exhibits the cities defined by the CCA with D^*=4560 and ℓ=3, as well as the discrepancy between the urban areas belonging to MSA/CMSA and CCA.

4. Conclusions

We analyzed the allometric scaling behavior of the NTL as a function of the population of the US cities. Our results corroborate previous works showing that the scaling behaviors of urban indicators with population can be substantially different for distinct definitions of city boundaries. Precisely, using the MSA/CMSA definition, we found a sublinear allometric scaling exponent $\beta_{MSA/CMSA} = 0.89 \pm 002$. Applying the CCA, we found an exponent $\beta_{CCA} = 1.01 \pm 0.02$ which indicates an isometric relation between the light pollution and the population of the US urban agglomerations, in clear contrast with the exponent obtained using the MSA/CMSA. Considering the consistency of the CCA definition in terms of the

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

extensivity between land population and area of their generated clusters, as demonstrated in previous studies for other urban indicators [15], we come to the conclusion that the proportionality between light pollution and population is indeed correct, as intuitively expected [37]. Under this framework and without loss of generality, it is therefore plausible to utilize NTL as a surrogate for city population in future studies.

The isometric relation between NTL and population of the US urban agglomeration, obtained applying the CCA, imply that small and large cities are proportionally indistinguishable in terms of light pollution. In other words, there is no *economy of scale* or sublinearity concerning the NTL in US cities. Our result shows that a growth of the US cities will aggravate the light pollution and therefore the possible negative effects of the light pollution for the humans and the wildlife health.

Acknowledgments

We gratefully acknowledge CNPq, CAPES, FUNCAP and the National Institute of Science and Technology for Complex Systems in Brazil for financial support. We specially thank our colleagues and friends H. P. M. Melo and T. A. Amor for the help and the discussions.

References

[1] F. Falchi, P. Cinzano, D. Duriscoe, C.C.M. Kyba, C.D. Elvidge, K. Baugh, B.A. Portnov, N.A. Rybnikova, R. Furgoni, The new world atlas of artificial night sky brightness, Sci. Adv. 2 (6) (2016) e1600377. http://dx.doi.org/10.1126/sciadv.1600377.
[2] N.A. Kerenyi, E. Pandula, G. Feuer, Why the incidence of cancer is increasing: The role of light pollution, Med. Hypotheses 33 (2) (1990) 75. http://dx.doi.org/10.1016/0306-9877(90)90182-E.
[3] D.E. Blask, G.C. Brainard, R.T. Dauchy, J.P. Hanifin, L.K. Davidson, J.A. Krause, L.A. Sauer, M.A. Rivera-Bermudez, M.L. Dubocovich, S.A. Jasser, D.T. Lynch, M.D. Rollag, F. Zalatan, Melatonin-depleted blood from premenopausal women exposed to light at night stimulates growth of human breast cancer xenografts in nude rats, Cancer Res. 65 (23) (2005) 11174. http://dx.doi.org/10.1158/0008-5472.CAN-05-1945.
[4] R.J. Reiter, F. Gultekin, L.C. Manchester, D. Tan, Light pollution, melatonin suppression and cancer growth, J. Pineal Res. 40 (4) (2006) 357. http://dx.doi.org/10.1111/j.1600-079X.2006.00325.x.
[5] K.J. Navara, R.J. Nelson, The dark side of light at night: Physiological, epidemiological, and ecological consequences, J. Pineal Res. 43 (3) (2007) 215. http://dx.doi.org/10.1111/j.1600-079X.2007.00473.x.
[6] R.J. Reiter, D. Tan, A. Korkmaz, T.C. Erren, C. Piekarski, H. Tamura, L.C. Manchester, Light at night, chronodisruption, melatonin suppression, and cancer risk: A review, Crit. Rev. Oncog. 13 (4) (2007) 303. http://dx.doi.org/10.1615/CritRevOncog.v13.i4.30.
[7] R. Chepesiuk, Missing the dark: Health effects of light pollution, Environ. Health Perspect. 117 (1) (2009) A20. http://dx.doi.org/10.1289/ehp.117-a20.
[8] R. Salgado-Delgado, A. Tapia Osorio, N. Saderi, C. Escobar, Disruption of circadian rhythms: A crucial factor in the etiology of depression, Depress. Res. Treat. 2011 (839743) (2011) 1. http://dx.doi.org/10.1155/2011/839743.
[9] M. Aubé, J. Roby, M. Kocifaj, Evaluating potential spectral impacts of various artificial lights on melatonin suppression, photosynthesis, and star visibility, PLoS One 8 (7) (2013) 1. http://dx.doi.org/10.1371/journal.pone.0067798.
[10] United Nations, World's population increasingly urban with more than half living in urban areas, 2014. URL http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html (accessed: 01.06.17).
[11] L.M.A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G.B. West, Growth, innovation, scaling, and the pace of life in cities, Proc. Natl. Acad. Sci. USA 104 (17) (2007) 7301. http://dx.doi.org/10.1073/pnas.0610172104.
[12] L.M.A. Bettencourt, G.B. West, A unified theory of urban living, Nature 467 (7318) (2010) 912. http://dx.doi.org/10.1038/467912a.
[13] L.M.A. Bettencourt, J. Lobo, D. Strumsky, G.B. West, Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities, PLoS One 5 (11) (2010) 20. http://dx.doi.org/10.1371/journal.pone.0013541.
[14] H.P.M. Melo, A.A. Moreira, É. Batista, H.A. Makse, J.S. Andrade, Statistical signs of social influence on suicides, Sci. Rep. 4 (6239) (2014) 1. http://dx.doi.org/10.1038/srep06239.
[15] E.A. Oliveira, J.S. Andrade, H.A. Makse, Large cities are less green, Sci. Rep. 4 (4235) (2014) 1. http://dx.doi.org/10.1038/srep04235.
[16] L.G.A. Alves, H.V. Ribeiro, E.K. Lenzi, R.S. Mendes, Empirical analysis on the connection between power-law distributions and allometries for urban indicators, Physica A 409 (Supplement C) (2014) 175–182. http://dx.doi.org/10.1016/j.physa.2014.04.046.
[17] F.J. Antonio, S.J. Picoli, J.J.V. Teixeira, R.S. Mendes, Growth patterns and scaling laws governing aids epidemic in brazilian cities, PLoS One 9 (10) (2014) 1–6. http://dx.doi.org/10.1371/journal.pone.0111015.
[18] L.M.A. Bettencourt, J. Lobo, Urban scaling in Europe, J. Roy. Soc. Interface 13 (116) (2016) 1. http://dx.doi.org/10.1098/rsif.2016.0005.
[19] C. Caminha, V. Furtado, T.H.C. Pequeno, C. Ponte, H.P.M. Melo, E.A. Oliveira, J.S. Andrade, Human mobility in large cities as a proxy for crime, PLoS One 12 (2) (2017) e0171609. http://dx.doi.org/10.1371/journal.pone.0171609.
[20] F.J. Antonio, A.S. Itami, S.J. Picoli, J.J.V. Teixeira, R.S. Mendes, Spatial patterns of dengue cases in brazil, PLoS One 12 (7) (2017) 1–16. http://dx.doi.org/10.1371/journal.pone.0180715.
[21] C.A. Kennedy, I. Stewart, A. Facchini, I. Cersosimo, R. Mele, B. Chen, M. Uda, A. Kansal, A. Chiu, K. Kim, C. Dubeux, E.L. La Rovere, B. Cunha, S. Pincetl, J. Keirstead, S. Barles, S. Pusaka, J. Gunawan, M. Adegbile, M. Nazariha, S. Hoque, P.J. Marcotullio, F.G. Otharán, T. Genena, N. Ibrahim, R. Farooqui, G. Cervantes, A.D. Sahin, Energy and material flows of megacities, Proc. Natl. Acad. Sci. USA 112 (19) (2015) 5985–5990. http://dx.doi.org/10.1073/pnas.1504315112.
[22] A. Facchini, C. Kennedy, I. Stewart, R. Mele, The energy metabolism of megacities, Appl. Energy 186 (2) (2017) 86–95. http://dx.doi.org/10.1016/j.apenergy.2016.09.025.
[23] H.A. Makse, S. Havlin, H.E. Stanley, Modelling urban growth patterns, Nature 377 (6550) (1995) 608. http://dx.doi.org/10.1038/377608a0.
[24] H.A. Makse, J.S. Andrade, M. Batty, S. Havlin, H.E. Stanley, Modeling urban growth patterns with correlated percolation, Phys. Rev. E 58 (1998) 7054–7062. http://dx.doi.org/10.1103/PhysRevE.58.7054.
[25] H.D. Rozenfeld, D. Rybski, J.S. Andrade, M. Batty, H.E. Stanley, H.A. Makse, Laws of population growth, Proc. Natl. Acad. Sci. USA 105 (48) (2008) 18702. http://dx.doi.org/10.1073/pnas.0807435105.
[26] H.D. Rozenfeld, D. Rybski, X. Gabaix, H.A. Makse, The area and population of cities: New insights from a different perspective on cities, Amer. Econ. Rev. 101 (5) (2011) 2205. http://dx.doi.org/10.1257/aer.101.5.2205.
[27] US Census Bureau, Cartographic Boundary Files, 2014. URL http://www.census.gov (accessed: 01.06.17).
[28] E. Doxsey-Whitfield, K. MacManus, S.B. Adamo, L. Pistolesi, J. Squires, O. Borkovska, S.R. Baptista, Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4, Pap. Appl. Geogr. 1 (3) (2015) 226. http://dx.doi.org/10.1080/23754931.2015.1014272.

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

[29] Socioeconomic data and application center (SEDAC), Gridded Population of the World, Version 4 (GPWv4), 2016. URL http://sedac.ciesin.columbia.edu (accessed: 01.06.17).
[30] S. Mills, S. Weiss, C. Liang, VIIRS day/night band (DNB) stray light characterization and correction, SPIE Opt. Eng. Appl. 8866 (88661P) (2013) 1. http://dx.doi.org/10.1117/12.2023107.
[31] National Centers for Environmental Information (NCEI), Visible Infrared Imaging Radiometer Suite (VIIRS), 2016. URL http://www.ngdc.noaa.gov/eog/viirs.html (accessed: 01.06.17).
[32] National Aeronautics and Space Administration (NASA), Visible Infrared Imaging Radiometer Suite (VIIRS), 2016. URL http://npp.gsfc.nasa.gov/viirs.html (accessed: 01.06.17).
[33] M. Shimrat, Algorithm 112: Position of point relative to polygon, Commun. ACM 5 (8) (1962) 434. http://dx.doi.org/10.1145/368637.368653.
[34] D.C. Montgomery, E.A. Peck, G.G. Vining, Introduction to Linear Regression Analysis, John Wiley & Sons, 2015.
[35] E.A. Nadaraya, On estimating regression, Theory Probab. Appl. 9 (1) (1964) 141.
[36] G.S. Watson, Smooth regression analysis, Sankhyā Ser. A (1964) 359.
[37] X. Li, X. Wang, J. Zhang, L. Wu, Allometric scaling, size distribution and pattern formation of natural cities, Palgrave Commun. 1 (15017) (2015) 1. http://dx.doi.org/10.1057/palcomms.2015.17.

Please cite this article in press as: F.G. Operti, et al., The light pollution as a surrogate for urban population of the US cities, Physica A (2017), https://doi.org/10.1016/j.physa.2017.11.039.

PLOS ONE

RESEARCH ARTICLE

Dynamics in the Fitness-Income plane: Brazilian states vs World countries

Felipe G. Operti[1,2], Emanuele Pugliese[3], José S. Andrade Jr.[1], Luciano Pietronero[2,3,4], Andrea Gabrielli[3]*

1 Departamento de Física, Universidade Federal do Ceará, 60451-970, Fortaleza, Ceará, Brazil, **2** Department of Physics, Sapienza University of Rome, 00185, Rome, Italy, **3** Istituto dei Sistemi Complessi (ISC) - CNR, UoS Sapienza, Dipartimento di Fisica, Università Sapienza, P.le Aldo Moro 5, 00185 - Rome, Italy, **4** International Finance Corporation, World Bank Group, 1818 H Street, 20433 - NW Washington DC - United States of America

* andrea.gabrielli@roma1.infn.it

OPEN ACCESS

Citation: Operti FG, Pugliese E, Andrade JS, Jr., Pietronero L, Gabrielli A (2018) Dynamics in the Fitness-Income plane: Brazilian states vs World countries. PLoS ONE 13(6): e0197616. https://doi.org/10.1371/journal.pone.0197616

Editor: Tobias Preis, University of Warwick, UNITED KINGDOM

Received: December 12, 2017

Accepted: May 4, 2018

Published: June 6, 2018

Data Availability Statement: Data are available from Dataviva. Data are provided by Ministério do Trabalho e Previdência Social (MTPS) – the Brazilian Ministry of Labor and Social Security – and Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC) – the Brazilian Ministry of Development, Industry and Foreign Trade. The data are available and can be used freely by anyone, without restrictions or control mechanisms (more informations at http://www.dataviva.info/en/help/). The data can be downloaded directly at http://www.dataviva.info/en/data/ or through the API at https://github.com/DataViva/dataviva-site/wiki.

Abstract

In this paper we introduce a novel algorithm, called *Exogenous Fitness*, to calculate the Fitness of subnational entities and we apply it to the states of Brazil. In the last decade, several indices were introduced to measure the competitiveness of countries by looking at the complexity of their export basket. Tacchella *et al* (2012) developed a non-monetary metric called Fitness. In this paper, after an overview about Brazil as a whole and the comparison with the other BRIC countries, we introduce a new methodology based on the Fitness algorithm, called Exogenous Fitness. Combining the results with the Gross Domestic Product *per capita* (GDP_p), we look at the dynamics of the Brazilian states in the Fitness-Income plane. Two regimes are distinguishable: one with high predictability and the other with low predictability, showing a deep analogy with the heterogeneous dynamics of the World countries. Furthermore, we compare the ranking of the Brazilian states according to the Exogenous Fitness with the ranking obtained through two other techniques, namely *Endogenous Fitness* and *Economic Complexity Index*.

Introduction

Large countries are often characterized by a strong internal heterogeneity between richer regions and poorer hierarchical regions. Just think to the difference between the GDP *per capita* (GDP_p) of the states of New York and Mississippi in the US [1], or the difference between the states of Kerala and Bihar in India [2], or between the unexplored forest of Amazon and the modern state of São Paulo in Brazil [3]. While the recent literature on Economic Complexity focused on countries [4–7], we believe that there are two very strong reasons to extend the scope of the analysis to the subnational level.

The first reason is purely academic. Indeed, sharp differences in economic outcomes in a uniform institutional area—with common cultural background and free movement of workers—are both a theoretical puzzle for traditional economics and an empirical opportunity for the Economic Complexity field. Indeed, the analysis of subnational entities competing on an even

Funding: FGO and JSAJ have been funded by the Brazilian Agency: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (www.capes.gov.br). JSAJ also has been funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (www.cnpq.br), Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (www.funcap.ce.gov.br), and National Institute of Science and Technology for Complex Systems (www.cbpf.br/inct-sc). EP, AG, and LP have been funded by the Italian PNR project CRISIS-Lab.

Competing interests: The authors have declared that no competing interests exist.

playing field is the perfect experimental setup to identify the role of organizational and technical capabilities with respect to more traditional economic factors of analysis. In this paper we will analyze the case of Brazil, to see if the capabilities driven dynamics of a country is replicated at a smaller scale.

The second reason is to improve economic forecasting. Indeed Economic Complexity has been proved to be very effective in forecasting the economic performances of countries [8]. An understanding of subnational entities could give however more accuracy and more detail. It is clear for example that future GDP_p growth of Brazil will depend not only on further growth of the Southern industrial core, but on the convergence of the other regions. This is crucial both to correctly forecast aggregate Brazil GDP_p growth and to address the vast internal inequality of the country.

Brazil or, officially, the *Federative Republic of Brazil* is the ninth World economy in the GDP ranking of the year 2015 [9]. Its population is equivalent to 2.81% of the total World population [10] and its large area (8.515.767,049 km^2), divided by its twenty-seven Federative Units [11], make it the fifth largest country of the World [12]. The political and administrative organization of Brazil is hierarchically organized in a sequence of geopolitical structures: Union, states, Federal District and Counties. Each one is autonomous and organized according with the division of powers: legislative, executive, and judiciary. Due to the deep inequalities, but also for the good perspectives of growth, Brazil and the others Latin American countries were often a focus of economic development analysis during the last century [13–17].

Economists usually focus on monetary based indices to analyze economies such as the GDP. However, GDP alone, as shown by different studies [4–7, 18], does not provide deep information about the perspective of growth and development of World countries. Several studies tried to gain information on the unobservable characteristics of countries by looking at stock indices to exploit the "wisdom of the crowd" [19, 20]. In order to gain a direct measure of the country capabilities, the last decade has been marked by a line of research of new indices inspired by the science of complex systems, able to better describe and explain the large scale World economy [4–7, 21–23] and to estimate global and regional inequalities [24, 25].

In this respect, different authors recently introduced two indices: Economy Complexity Index (ECI) [4] and Fitness [5]. Furthermore, Cristelli *et al* [7], through a novel method called *Selective Predictability Scheme* (SPS), showed that the comparison between GDP_p and Fitness provides a highly performing forecasting tool for several countries.

In this paper, we first present an overview of Brazil as a whole from the point of view of the Economic Complexity approach. In this context we compare its export basket and its Fitness with the ones of the BRIC group of countries (Brazil, Russia, India, and China) [26].

Then, we focus on the comparative study of the economies of the single Brazilian states. Based on the "classical" Fitness algorithm, we introduce a new methodology, called Exogenous Fitness, able to measure the Fitness of subnational entities, and we apply it to the states of Brazil. In analogy with what was proposed in [7], we analyze the coevolution of GDP_p and Fitness studying in this way the predictability of the economic growth of the Brazilian states.

Furthermore, we compare the Exogenous Fitness with: (i) the (Endogenous) Fitness -i.e, the natural application of the "classical" Fitness algorithm to the subnational entities of a country-; (ii) the results published by the *Dataviva* platform (an application of the ECI algorithm) [27].

The paper is structured as follows: first we introduce the methods and we provide an overview about Brazil. Then, we show the results of the Exogenous Fitness applied to the Brazilian states and the comparisons with the other techniques. Finally, we conclude with a general discussion about the implications of the results with respect to both points of view of scientific community and policy makers. In the Appendix A, we describe in detail the used database.

Methods

In this section we describe the algorithms and methods involved in the calculation of the states and countries Fitness coupled to the Complexity of exported products.

Revealed Comparative Advantage (RCA)

The Revealed Comparative Advantage (RCA) [28] is a quantitative criterion to assess the relative advantages of a country, or, in this case, of a Brazilian state, in the export of certain products compared to the average export of those products. Defining q_{sp} as the flow of the export (in US dollars) of the product p by the state s (see the section *Database* for the data origin), the RCA is defined as:

$$RCA_{sp} = \frac{\frac{q_{sp}}{\sum_{p'} q_{sp'}}}{\frac{\sum_{s'} q_{s'p}}{\sum_{s'p'} q_{s'p'}}}. \quad (1)$$

Therefore, it is the ratio between the share of the export of product p with respect to the total export of State s divided by the share of the export of product p with respect to the total Brazilian export.

From the calculation of the RCA for each state-product pair, we build the binary state-product matrix M_{sp}. We consider the state s an exporter of a product p, if $RCA_{sp} \geq 1$ and, consequently, we set $M_{sp} = 1$. On the contrary, if $RCA_{sp} < 1$, we set $M_{sp} = 0$.

An analogous criterion is used to define the World countries-products matrix M_{cp} (see the section *Database*). This binary matrix shows which country has a comparative advantage in a certain product with respect to the World average [29, 30].

(Endogenous) Fitness

Recently, different studies have shown the economic relevance of the diversification of the export basket for the competitiveness of a country [4, 6]. The matrix M shows a substantial nested structure highlighted by a strong triangularity, which can be interpreted in the following way: each country approximately exports all the possible products it has the capabilities to produce [5].

Here, considering the geographic size of Brazil and its federal structure, we assume that the same concept is also valid to understand the development and growth of its states. In this framework, we apply the Fitness algorithm to the states-products matrix of elements M_{sp} above defined [5], a statistical approach based on non linear maps coupling Fitness of states and Complexity of Products, to compare Brazilian states. The (Endogenous) Fitness algorithm is defined by the following iterative equations [6]:

$$\begin{cases} \tilde{F}_s^{(n)} = \Sigma_p M_{sp} Q_p^{B\ (n-1)} \\ \tilde{Q}_p^{B\ (n)} = \dfrac{1}{\sum_s M_{sp} \frac{1}{F_s^{(n-1)}}} \end{cases} \rightarrow \begin{cases} F_s^{(n)} = \dfrac{\tilde{F}_s^{(n)}}{\langle \tilde{F}_s^{(n)} \rangle_s} \\ Q_p^{B\ (n)} = \dfrac{\tilde{Q}_p^{B\ (n)}}{\langle \tilde{Q}_p^{B\ (n)} \rangle_p}. \end{cases} \quad (2)$$

The elements M_{sp} are the elements of the previously discussed binary states-products matrix. $\tilde{F}_s^{(n)}$ and $\tilde{Q}_s^{B\ (n)}$ are intermediate variables which are subsequently normalized at each iteration.

The initial conditions satisfy the relations: $\tilde{F}_s^{(0)} = C$ and $\tilde{Q}_p^{B\ (0)} = C$, where we assume $C = 1$ for each state s and for each product p [6].

At each iteration of the algorithm, the Fitness of each state is proportional to the sum of its exported products weighted by their Complexity stressing the importance of having at the same time both a diversified export basket and the most complex possible products in it. The formula for the Complexity of a product is motivated by the following argument: the more the exporters of a product and the smaller their Fitness, the less its expected from the Complexity. In this manner, a state with low Fitness abruptly influences the Complexity of all the products it exports [6]. Therefore, an highly Complex product is made only by few countries/states with high Fitness, while a little Complex product can be made by all the countries/states, both with high and low Fitness. The stability and robustness of the algorithm has been studied in [6, 31] and the Fitness ranking of the states and the Complexity ranking of the products is unambiguously defined after a large enough number of iterations.

Fig 1 shows the matrix M_{sp} of the year 2015, by ordering the states according to the Fitness (the upper the higher complexity), and the products according to the Complexity (the more right the higher the complexity).

In Fig 2 we show the products *spectroscopy* [32] of the years 2005 (dashed lines) and 2015 (filled colors) for few Brazilian states such as: São Paulo, Paraná, Ceará, and Roraima. The spectroscopy is a graphic representation of the export volume (in US Dollars) of a state for each product with $M_{sp} = 1$ ordered at increasing Complexity from left to right [32]. We subsequently group the products (10 for bin) and we summed the export volumes of each product inside each bin. The spectroscopy allows to compare the diversification and the Complexity of the exportation of the states. The figure shows the spectroscopy of high Fitness states such as São Paulo (diversified all along the Complexity spectrum) and Paraná (with a clear peak on medium-high Complexity products), a middle rank state such as Ceará and a low Fitness state such as Roraima (with few low Complexity exports). From the figure, it emerges that a very developed state such as São Paulo has a high flow of exports for a very diversified number of products with a bias towards the high Complexity ones. Paraná has a high peak in several complex products, while Roraima has only one peak in the less complex products. Ceará is a middle ground between the two.

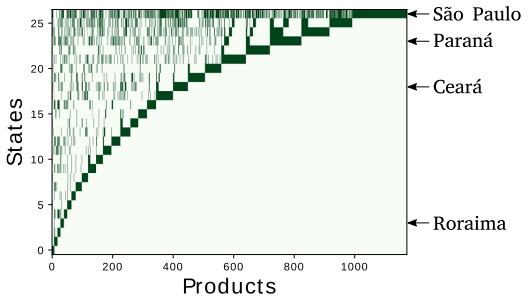

Fig 1. The binary matrix M_{sp} of the year 2015. Each row of the matrix represents a Brazilian state. States are ordered in terms of their Fitness from the smallest value (row 0) to the largest one (row 26). Analogously columns represent Products ordered in terms of their Complexity from the smallest value (column 0) to the largest one (column 1172). The matrix elements M_{sp} are drawn in dark green and the others in white. In the figure we highlight high Fitness states such as São Paulo and Paraná, a middle rank State such as Ceará and a low Fitness state such as Roraima.

https://doi.org/10.1371/journal.pone.0197616.g001

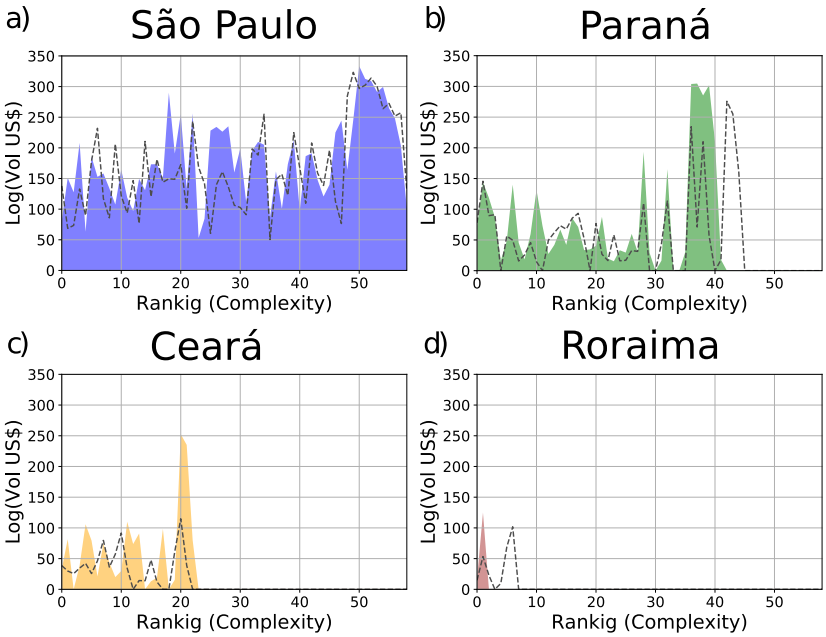

Fig 2. Products spectroscopy of the years 2005 (dotted lines) and 2015 (filled colors) of the states: a) São Paulo, b) Paraná, c) Ceará, and d) Roraima. The figures show the export volume (in US Dollars) of those states for each product with $M_{cp} = 1$ ordered according to their Complexity. Products are grouped in bins of 10 and the export volume in each bin are summed up.

https://doi.org/10.1371/journal.pone.0197616.g002

Exogenous Fitness

Here, we define the new Exogenous Fitness algorithm, an innovative method to calculate the Fitness of subnational entities of a country grounded on the measure of the products Complexity from the World-wide trade network. Exogenous Fitness is a coherent extension of the "classical" Fitness algorithm [6], with the assumption of an obvious concept: products have an intrinsic Complexity, reflected by the trade on the global World scale by all countries, while the trade from the regions of a single country may not represent well such intrinsic Complexity as it can be affected by local biases. In particular if we consider only Brazil to define the Complexity of the exported products, we can introduce local economic biases in its measure related to the peculiar features of Brazil economy. Indeed, as shown in Fig 1, there is a big range of products made only by few states that make the measure of Complexity very inaccurate. From this observation, it is natural to use as the best measure of Complexity of products the ones Q_p^W extracted from the Fitness algorithm applied to the trade of goods of all World countries, i.e. we take:

$$Q_p^W \equiv Q_p^B \equiv Q_p. \tag{3}$$

Indeed, the Complexities of the products obtained applying the Endogenous Fitness to the

World countries (Q_p^W) can be considered the same of the Complexities of the products inside Brazil (Q_p^B) and, therefore, we simply define them as Q_p.

Therefore, the algorithm consists of two steps:

1. We apply the (Endogenous) Fitness (Eq 2) to the World countries, as previously done in [5–7]. The criterion adopted to determine if a country c is a "good" exporter of a given product p is again based on the RCA extended to all World countries: we set $M_{cp} = 1$ if $RCA_{cp} \geq 1$ and $M_{cp} = 0$ otherwise (see the section *Database* for the source of the data). Applying the (Endogenous) Fitness algorithm to the matrix M_{cp}, after a sufficiently large number of iterations the algorithm converges to the fixed point so that, we obtain the respective Fitness F_c for each country and the Complexity Q_p^W for each product.

2. From the assumption Eq 3, we use as Complexity of the products exported by Brazilian states Q_p the values obtained by the Fitness algorithm applied to the export of all World countries. Therefore, we use the information in the matrix M_{sp} and the product Complexity Q_p to calculate the Fitness of the Brazilian states through the following formula:

$$\begin{cases} \tilde{F}_s = \Sigma_p M_{sp} Q_p \\ F_s = \dfrac{\tilde{F}_s}{\langle \tilde{F}_s \rangle_s}. \end{cases} \tag{4}$$

The relevance of developing the Exogenous Fitness measure is two folds. First of all, using world wide data we extract all the information to compute the Complexity of products to better compute the Fitness of states. Since the algorithm works by exploiting differences of capabilities, using world wide data we gain additional information related to the export baskets of countries with a wider range of Fitness and capabilities. Of course we still expect the two measures to be highly correlated in rank, in particular for a country like Brazil that contains such a vast array of development levels. As we will see in section *Comparison with other techniques*, this is indeed the case. The second reason is that the Exogenous Fitness allows to have for states Fitness values comparable with those of countries. Indeed, while the ranking between Exogenous and Endogenous Fitness are highly correlated, their actual values and distributions are vastly different. As detailed explained in the paper [33], while the ranking for the Fitness measure is always well defined, the shape of the matrix directly affects the convergence properties of the algorithm to a polarized distribution. Employing the Exogenous Fitness method we have smoothly changing values that allows for the forecasting exercises of section *Results*.

Overview of Brazil

First, we analyze Brazil as a whole applying the (Endogenous) Fitness to World countries in the time interval from 1995 to 2015. In Fig 3 we show the matrix M_{cp} of the World countries of year 2015 obtained by ordering the countries according to the Fitness and the products according to the Complexity. In that year, Brazil is ranked in the 44th/147 position (equivalent to the raw 103 in Fig 3).

In Fig 4, we show the dynamics of the World countries in the Fitness-Income plane emphasizing the BRIC countries (Brazil in green, Russia in blue, India in orange, and China in red). The figure shows that India and China have in 1995 lower values of GDP_p than Brazil and Russia, but higher values of Fitness. According with [7], this difference justifies the dynamics in the plane of the four countries for the next years. Indeed, India and China continued their

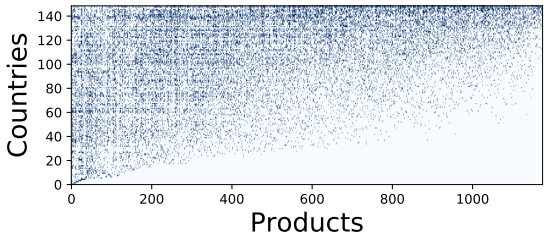

Fig 3. The binary matrix M_{cp} of the year 2015. The rows of the matrix represent the World countries ordered according to their Fitness with row 0 for the country with the lowest Fitness and row 147 for the one with the highest Fitness. Analogously columns represent Products ordered in terms of their Complexity from the lowest one at column 0 to the highest one at column 1174. The elements $M_{cp} = 1$ are represented as blue dots.

https://doi.org/10.1371/journal.pone.0197616.g003

economic growth during the following years, while Russia and Brazil entered a period of recession [34].

In order to zoom on the differences among the dynamics of the BRIC countries, we analyze the variation of the Fitness of such countries during the interval from 2003 and 2013. The variation of the Fitness can have two different causes: (i) changes in the export basket, (ii) changes in the products Complexity. We can decompose the variation of Fitness [35] as:

$$\Delta\tilde{F}_c = \tilde{F}_c(t_1) - \tilde{F}_c(t_0) = \sum_p M_{cp}(t_1)Q_p(t_1) - \sum_p M_{cp}(t_0)Q_p(t_0) = \\ = \sum_p \Delta M_{cp}\frac{Q_p(t_1) + Q_p(t_0)}{2} + \sum_p \Delta Q_p \frac{M_{cp}(t_1) + M_{cp}(t_0)}{2}. \quad (5)$$

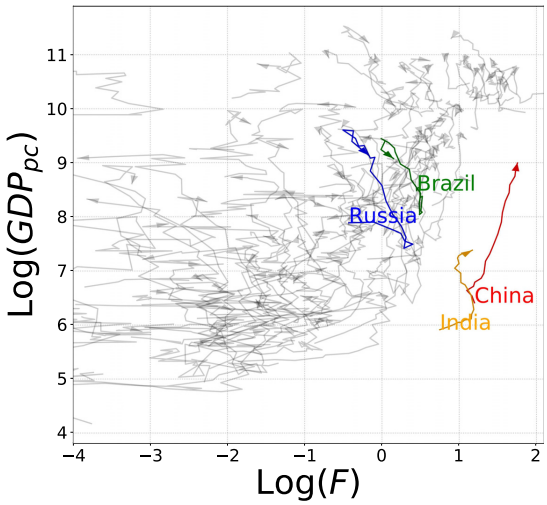

Fig 4. Dynamics of the World countries in the Fitness-Income plane. The figure shows the dynamics (from the year 1995 to the year 2015) of World countries in the Fitness-Income plane in logarithmic scale. We emphasize the BRIC countries: Brazil in green, Russia in blue, China in red, and India in orange.

https://doi.org/10.1371/journal.pone.0197616.g004

Table 1. Fitness variation from 2003 to 2013 of BRIC countries.

	Variation due to changes in the export basket	Variation due to changes in the products Complexities
Brazil	-43%	-6%
Russia	-37%	-21%
China	+32%	+18%
India	-18%	+2%

https://doi.org/10.1371/journal.pone.0197616.t001

where we have indicated with $\Delta X = X(t_1) - X(t_0)$ for a generic quantity X. The first term in the last step of the equation is the contribution to $\Delta\tilde{F}_c$ due to the variation in the export basket, while the second one is the term due to variation of products Complexities. In Table 1, we show both the percentage variations due to the two terms. The results show a deep decrease of both terms for Russia and we can see how the loss of competitiveness of Brazil is mostly due to the drop of products that were previously exported, and not so much related to the change in complexity of those products. In contrast China has increased its export basket and the Complexity of the exported products. Instead, India in 2013 exports more complex products, but has decreased its exports diversification.

Furthermore, we show in Fig 5 the products spectroscopy [32] for the BRIC countries of the year 2005 (dotted lines) and 2015 (filled colors). The figure shows that Brazil and Russia

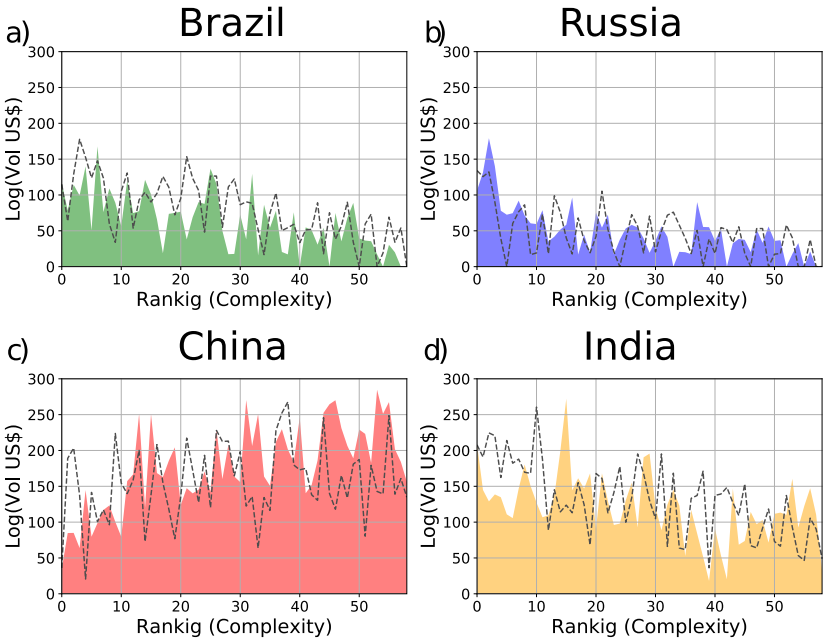

Fig 5. Products spectroscopy of the years 2005 (dotted lines) and 2015 (filled colors) of the countries: a) Brazil, b) Russia, c) China, and d) India. The figures show the export volume (in US Dollars) of those states for each product with $M_{cp} = 1$ ordered in terms of their Complexity. The products have been grouped (10 for bin) and the export volumes of each product inside each bin have been summed.

https://doi.org/10.1371/journal.pone.0197616.g005

have a high exportation only of simple products, while India and China have a high exportation of complex products.

Therefore, Figs 4 and 5 and Table 1, show that China and India both have a diversified export basket and export complex products. Such factors determine a high Fitness and consequently a growth of the GDP_p in the subsequent years. On the contrary Brazil and Russia export simple products with a consequently low Fitness so that these countries entered a recession period [34].

In the next section we show the results of a deepened analysis of the internal economy of Brazil through the application of the Exogenous Fitness to the Brazilian states.

Results

We applied the Exogenous Fitness algorithm to the Brazilian states in the time interval from 2000 to 2015 obtaining for each year both well-defined values of Fitness for each Brazilian state, and the ranking of states in terms of their Fitness (shown in Fig 6).

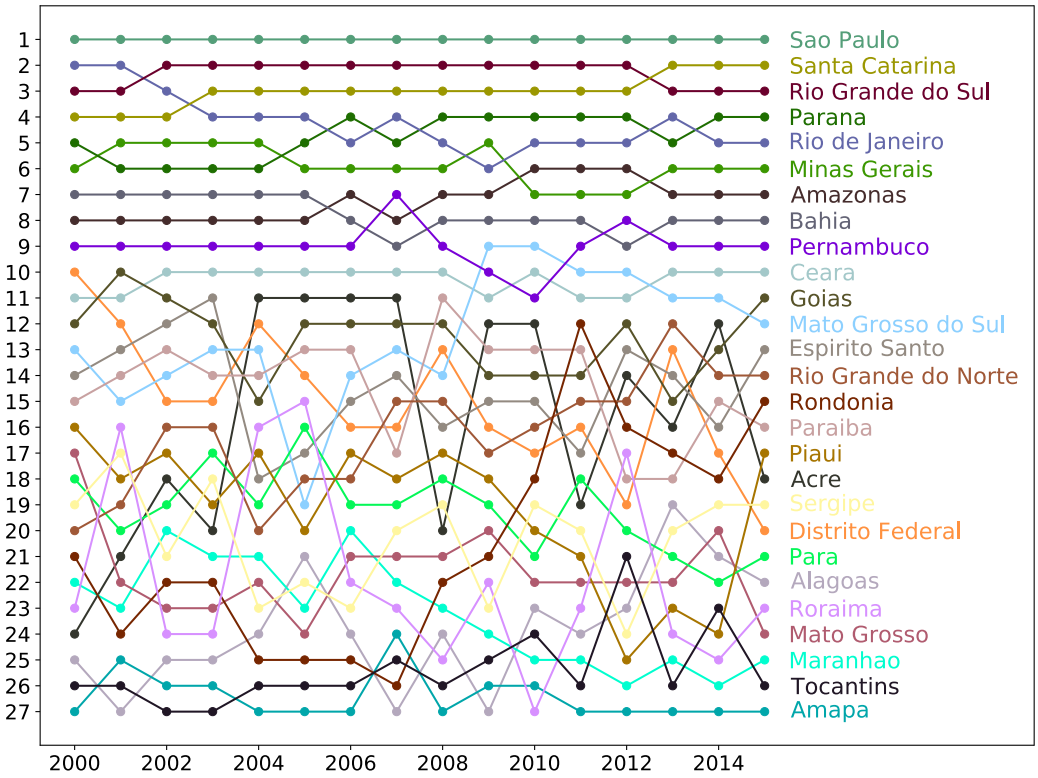

Fig 6. Time evolution of the ranking of Brazilian states according to the Exogenous Fitness algorithm. The figure shows the time evolution of the ranking of the Brazilian states according to the Fitness obtained through the Exogenous Fitness algorithm applied to the time interval 2000-2015.

https://doi.org/10.1371/journal.pone.0197616.g006

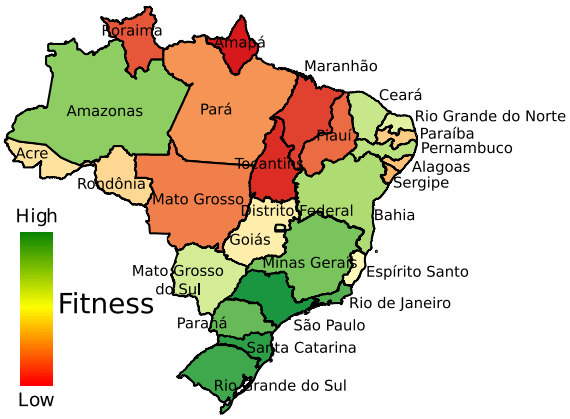

Fig 7. Fitness map of the Brazilian states. The colors in the map vary from green (high Fitness) to red (low Fitness) and they show the differences of the Fitness among the Brazilian states.

https://doi.org/10.1371/journal.pone.0197616.g007

We show in Fig 7 a map of Brazil where each state is colored according to its Fitness. From the figure, it emerges Southern states have larger Fitness, and therefore have a better economic development, than Northern states. This result is in agreement with other monetary and non-monetary indices such as the Human Development Index (HDI) and the GDP [27].

Furthermore, we show in Table 2 the variation from 2003 to 2013 of the Fitness ($\Delta\tilde{F}_s$) for several states such as: São Paulo (1st in Fitness ranking of year 2013), Paraná (5th in Fitness ranking of year 2013), Ceará (10th in Fitness ranking of year 2013), and Roraima (24th in Fitness ranking of year 2013). From both Fig 2 and Table 2 we observe that São Paulo has a diversified export basket with high peaks in complex products and, at the same time, it increases both the export basket and the Complexity of the exported products in the considered time period. Paraná and Ceará, in contrast with the aggregate behavior of Brazil, in the same period grew in diversification becoming more competitive—even in the face of a minor decline in the complexity of their exported products. Roraima, on the contrarty, shows a deep decrease in the diversification.

As mentioned in the previous section, Fig 4 presents the dynamics of World countries in the Fitness-GDP$_p$ plane. It shows a high degree of heterogeneity of the dynamics of countries. Indeed, the plane can roughly be divided into two regions: one with an unpredictable "chaotic" regime of the evolution of countries, and the other with a predictable "laminar" regime. In order to overcome the limitations of linear regressions, Cristelli *et al* [7] proposed an

Table 2. Fitness variation from 2003 to 2013 of the states: São Paulo, Paraná, Ceará, and Roraima.

	Variation due to changes in the export basket	Variation due to changes in the products Complexities
São Paulo	+2%	+2%
Paraná	+59%	-7%
Ceará	+53%	-10%
Roraima	-37%	+6%

https://doi.org/10.1371/journal.pone.0197616.t002

innovative data-driven non-parametric prediction scheme called the *Selective Predictability Scheme* (SPS). It is inspired by the so-called *method of analogues* [36, 37] and through a *measure of concentration* it delimits predictability regions inside the Fitness-Income plane. The measure of concentration consists in dividing the plane into a grid and analyzing the time evolution of the distribution of countries inside each box with at least five countries inside.

In analogy with what has just been explained for World countries, in Fig 8a, we show the time evolution of the real GDP_p as a function of the Fitness (obtained implementing the Exogenous Fitness algorithm), for each Brazilian state in the period 2000-2015. The dotted black line in the figure shows the expected level of GDP_p given the level of Fitness and it is the result of the minimization of the Euclidean distance of the states from the line, weighted by the state GDP. From the figure emerges an heterogeneous dynamics similar to the dynamics of World countries that cannot be analyzed through a linear regression. Also the *measure of concentration* is not appropriate in this case. Indeed the reduced number of Brazilian states (27) compared with the number of World countries (146) makes this measure inappropriate for the internal analysis of Brazil. In order to have a significant number of cells with at least five states, the granularity of the grid should be too broad to analyze the evolution of the distribution.

Therefore, in order to validate the predictability of the dynamics of the states in the Fitness-Income plane, here we develop a novel intuitive method, the *measure of direction*. First of all let us fix the time window $[t_1, t_2]$ in which we want to study the evolution of each state in the plane $\log(Fitness) - \log(GDP_p)$. The time lag $\Delta = t_2 - t_1$ has to be taken large enough to get a sufficient noise reduction in the dynamics. We choose $t_1 = 2003$ and $t_2 = 2013$. Second, we divide the plane in a fine grid of 100×100 cells and we define two bandwidth; one for the x-axis, and the other for the y-axis. For each cell, we define around its centroid a threshold area of sides given by the two bandwidths. Then, for each cell k with at least three states at the time t_1 inside its threshold area, we computed the average dot product $\tilde{D}_k$:

$$\tilde{D}_k = \frac{2}{N(N-1)} \sum_{i<j}^{1,N} \hat{v}_i \cdot \hat{v}_j, \tag{6}$$

where $\hat{v}_i = \frac{\vec{v}_i}{v_i}$ where $\vec{v}_i = [log(F_i(t_2)) - log(F_i(t_1))]\hat{i} + [log(GDP_{p_i}(t_2)) - log(GDP_{p_i}(t_1))]\hat{j}$ and $\hat{i}$ and $\hat{j}$ are respectively the versors in the Fitness and GDP_p directions. N is the number of states with starting point inside the threshold area of cell k. The coefficient $\tilde{D}_k$ gives the average cosine among the versors of all states initially inside the threshold area of cell k and varies from $(-1, 1]$. It measures the dispersion of the directions of evolution in the plane in the time window $[t_1, t_2]$ of all states initially in the threshold area of cell k: when it is close to 1 all states initially in the threshold area of cell evolve in a coherent parallel way. The smaller is $\hat{D}_k$ the larger the dispersion of these trajectories. A color map of the coefficient $\tilde{D}$ in the different cells is shown in Fig 8b. From the figure it emerges that there is a region where the directions of evolution of the states tend to be parallel (showed in green) and a region where the directions of motion tend to be unevenly directed (showed in red). Increasing/decreasing the bandwidths and, therefore, the threshold area only changes the resolution of the image, but the two regions remain well-defined. In Fig 8b we used an x-axis bandwidth 0.86, and a y-axis bandwidth 0.38, providing an almost continuous variation of the colors map.

In order to investigate which is the main direction of the versors in the green region and the further directions in the red region, we divided the plane into a broader grid (10x10). For each cell we sum all the vectors inside it and then we calculate the versor of the sum vector. We show the result in Fig 8c. From one hand, from the figure we can observe a region where the states tend to evolve in the same direction (shown in green). Therefore, in this region, the

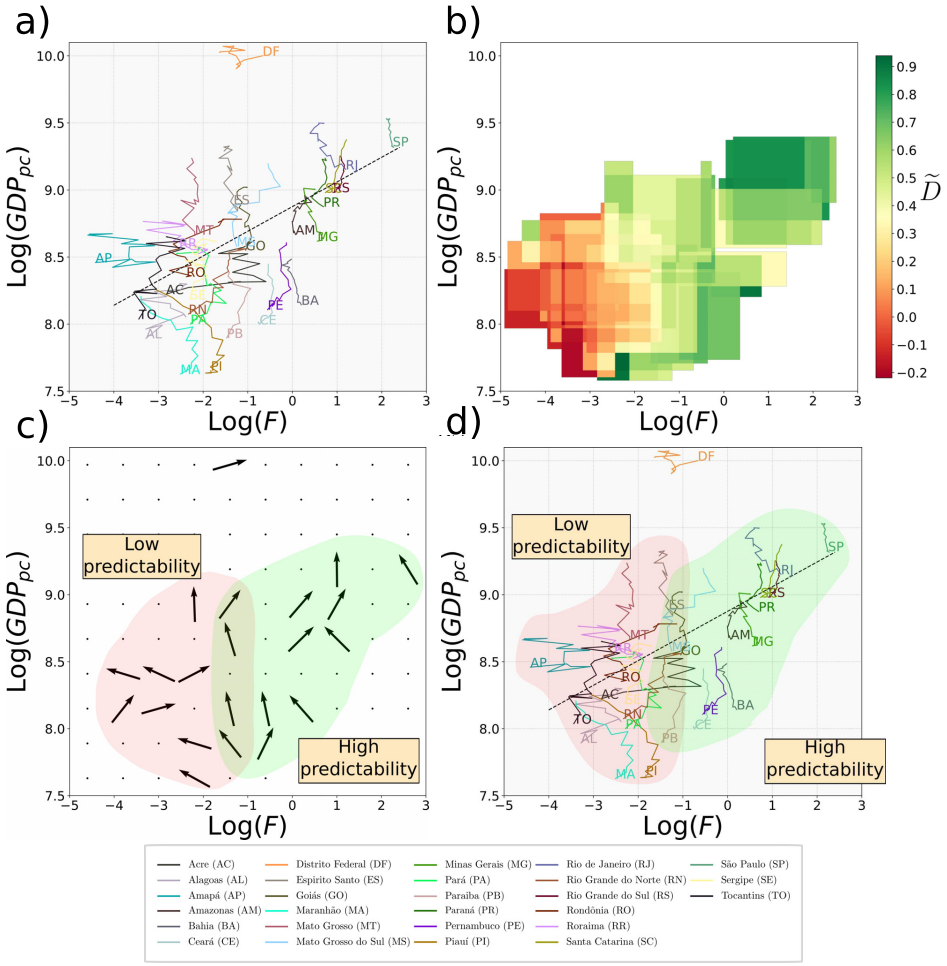

Fig 8. Dynamics of Brazilian states in the Fitness-Income plane. *a*) The figure shows the evolution (from 2000 to 2015) of the Brazilian states in the Fitness-Income plane in logarithmic scale. The dotted black line in the figure shows the expected level of GDP_p given the level of Fitness and it is the result of the minimization of the Euclidean distance of the states from the line, weighted by the states GDP. *b*) The figure shows the coefficient $\widetilde{D}$ calculated considering a time window from 2003 to 2013. The color varies from green (where the versors of evolution tend to be parallel), to red (where the versors tend to be unevenly directed). *c*) The figure shows a grid where for each cell we calculate the versor of the sum vector. From the figure two regions appear: the first one where the versors tend to be parallel in the direction of a high GDP_p (shown in green); and the second one where the versors tend to be unevenly directed (shown in red). Fig 8*b* and *c* together show that there is a region (green) of high predictability of motion in direction of a high GDP_p; and a region (red) of low predictability of motion. *d*) The figure shows the dynamics (from 2000 to 2015) of the Brazilian states in the Fitness-Income plane highlighting in green the states in the high predictability region and in red the states in the low predictability one.

https://doi.org/10.1371/journal.pone.0197616.g008

future evolution of countries is predictable with good confidence. On the other hand, another region (shown in red) can be detected where the versors tend to be unevenly directed. The dynamics of the states in this region is basically unpredictable. Furthermore, in the middle of the two, there is a region of transition, shown in the figure by the overlapping of the two colors.

Lastly, in Fig 8d we show the dynamics of the states in the Fitness-Income plane highlighting in green the states with high predictability of the motion and in red those with low predictability. From the figure emerges that states as Ceará, Pernambuco, and Bahia, despite having low values of GDP, are in a region of high Predictability and, therefore, they will probably continue to growth in the same direction. While for states as Acre, Tocantins, or Alagoas the dynamics is more chaotic and predictions are less reliable.

Comparison with other techniques

In this section we compare the results obtained implementing the Exogenous Fitness with the results of the Endogenous Fitness and the ones published by Dataviva [27] obtained by applying the Economic Complexity Index (ECI).

Exogenous Fitness and Endogenous Fitness

We apply the (Endogenous) Fitness algorithm to the Brazilian states in the time interval from 2000 to 2015 obtaining the time evolution of the ranking of the states according to such kind of Fitness (shown in Fig 9). Calculating the Spearman correlation coefficient between the ranking obtained through the Exogenous and the Endogenous Fitness for each year in the analyzed time interval, we obtain an average value $\tilde{\rho}_{ExEn} = 0.97$. This result shows a strong correlation between the rankings obtained through the two different Fitness algorithms.

The Endogenous Fitness algorithm provide us a well-defined annual ranking of the Brazilian states, but not well-defined quantitative values of Fitness and products Complexity. In fact, all Fitness values except one tend to zero. After a fairly high number of iterations, however, the ranking of states stabilizes, and there are no more changes of ranking among the states. This circumstance is already been studied [33] and it is due to the shape of the matrices M_{sp}. Indeed the *external area* (where $M_{sp} = 0$) is greater than the *internal area* (where almost all elements $M_{sp} = 1$) for each analyzed year.

Exogenous Fitness and ECI

In Fig 10 we show the time evolution (from 2002 to 2015) of the ranking of the Brazilian states according to ECI, directly downloaded by the Dataviva platform [27]. Therefore, in order to compare the ranking obtained through the Exogenous Fitness algorithm and the ECI algorithm, we calculate the annual Spearman correlation coefficient between the two rankings in the period 2002-2015, obtaining an average value $\tilde{\rho}_{ExECI} = -0.14$. This result shows an almost total absence of correlations between the two rankings, i.e. between the two algorithms.

Indeed, already from a qualitative point of view, ECI ranking seems to be unrealistic. For example, it ranks rich states in GDP, but also with high HDI [27], such as Santa Catarina or Paraná, in the last positions (respectively 26th and 24th position in 2015). Moreover, the state of Alagoas (last in HDI ranking of 2014 [27]) is unrealistically ranked in 4th position in the 2015.

In Fig 11, we show the map of Brazil where each state is colored according to its ECI. From the figure, it emerges that there is no geographic coherence among the ECI of the different states. For instance the figure shows that the state of Santa Catarina has a high ECI, but it is in the middle between the states of Rio Grande do Sul and Paraná that have a low ECI.

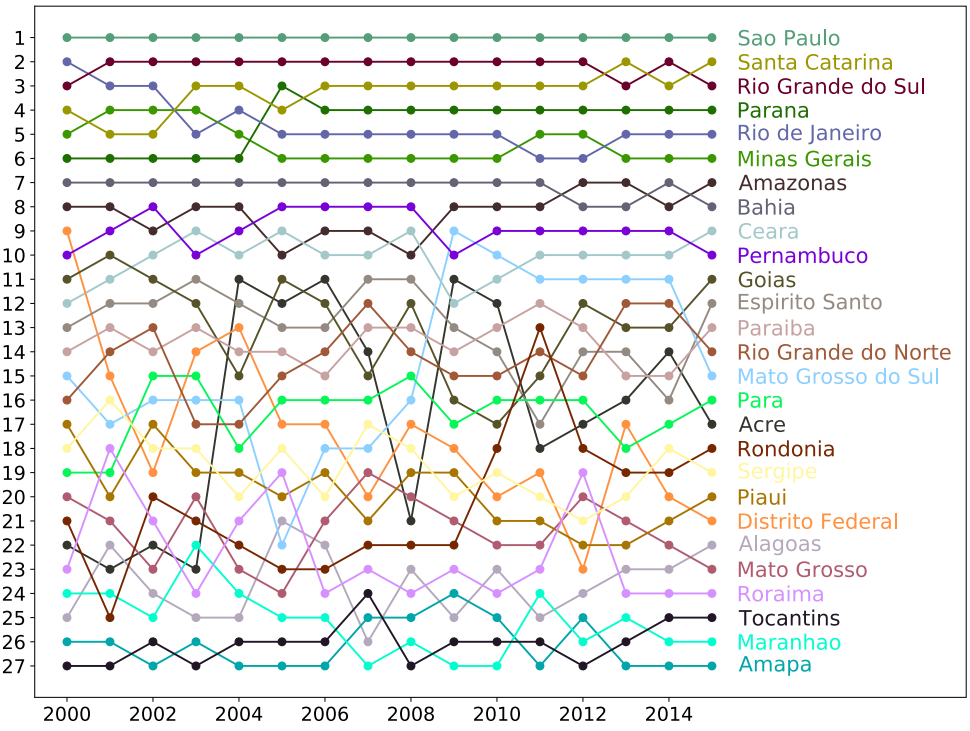

Fig 9. Time evolution of the ranking of Brazilian states according to the (Endogenous) Fitness algorithm. The figure shows the time evolution of the ranking of the Brazilian states in terms of the Fitness obtained through the (Endogenous) Fitness algorithm applied during the time interval 2000-2015.

https://doi.org/10.1371/journal.pone.0197616.g009

Furthermore, we show in Fig 12a the evolution of Brazilian states in the ECI-Income plane, where the income is in logarithmic scale. In Fig 12b, we show the coefficient $\tilde{D}$ above defined but applied to ECI instead to *log*(*Fitness*) and in Fig 12c the directions of motions. Differently from the results obtained through the application of the Exogenous Fitness (Fig 8), using the ECI index the dynamics of the states is unpredictable. Indeed, all the states except São Paulo and the Distrito Federal are concentrated in a small region of the plane and, therefore, totally indistinguishable.

Discussion

In this paper we first compared the dynamics of Brazil in the Fitness-Income plane with the other BRIC countries. In Fig 4, we observed that IC (India and China) countries, both with a high Fitness compared to the BR (Brazil and Russia) countries, grow in GDP_p for the entire analyzed time interval. Table 1 shows that IC improve the Complexity of export baskets in the analyzed time interval, and China even shows an improvement of the diversification. Instead, BR countries did not invested in diversification and in Complexity of the exported products

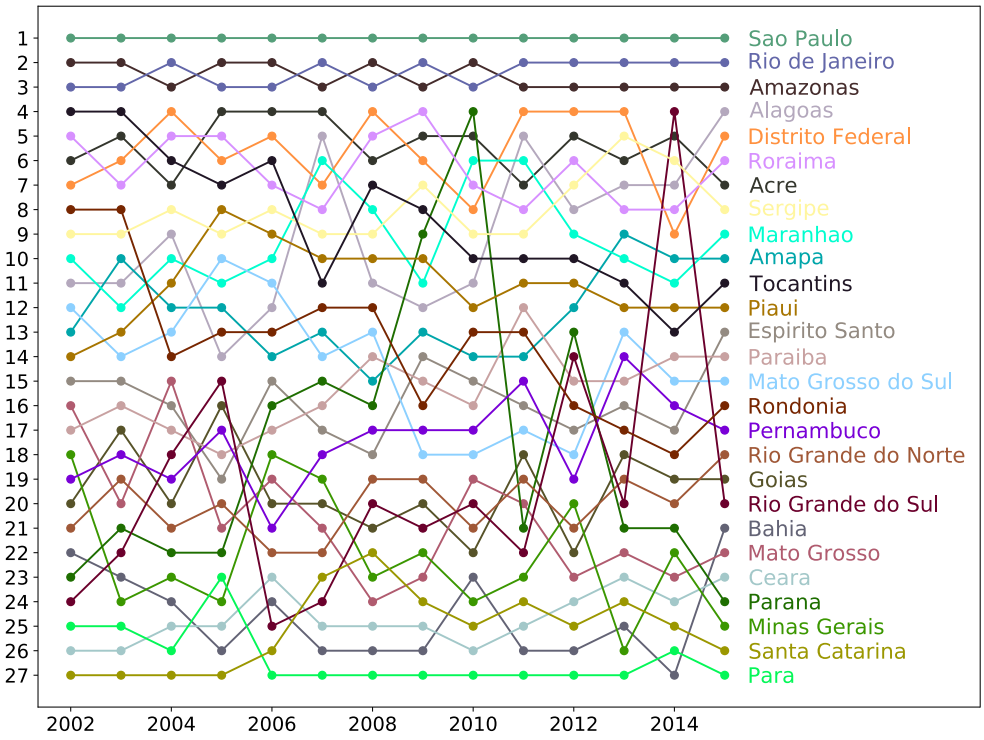

Fig 10. Time evolution of the ranking of Brazilian states according to the ECI algorithm. The figure shows the time evolution of the ranking of the Brazilian states during the period 2002-2015 in terms of the ECI, directly downloaded by the Dataviva platform [27].

https://doi.org/10.1371/journal.pone.0197616.g010

(as shown in Table 1). These results strengthen an hypothesis previously formulated in [7]: Fitness is the driving force behind growth.

In the second part of the paper, we introduced a new algorithm called "Exogenous Fitness" to calculate the Fitness of subnational entities and we applied it to the states of Brazil. The comparison between the Fitness and the GDP_p showed an heterogeneous dynamics of the Brazilian states in the Fitness-Income plane. Indeed, two regions are distinguishable in the plane: one with high predictability and the other with low predictability. Here, we have shown that economic forecasting is possible for those states in the high predictability region, while it is not for those in the low predictability region. As a consequence of this analysis Fitness seems to be the driving force behind growth. Indeed, the dynamics in the high predictability region is characterized by high values of Fitness, while high value of GDP_p is not a good signature of growth. The heterogeneous dynamics observed for the Brazilian states shows a strict analogy with the heterogeneous dynamics observed for the World countries [7]. Furthermore, by comparing the export "spectroscopy" of BRIC countries with the one of Brazilian states of São Paulo, Paraná, Ceará, and Roraima, and, comparing the variations of the Fitness, we observe that countries/states with diversified export baskets produce high complex products and grew in

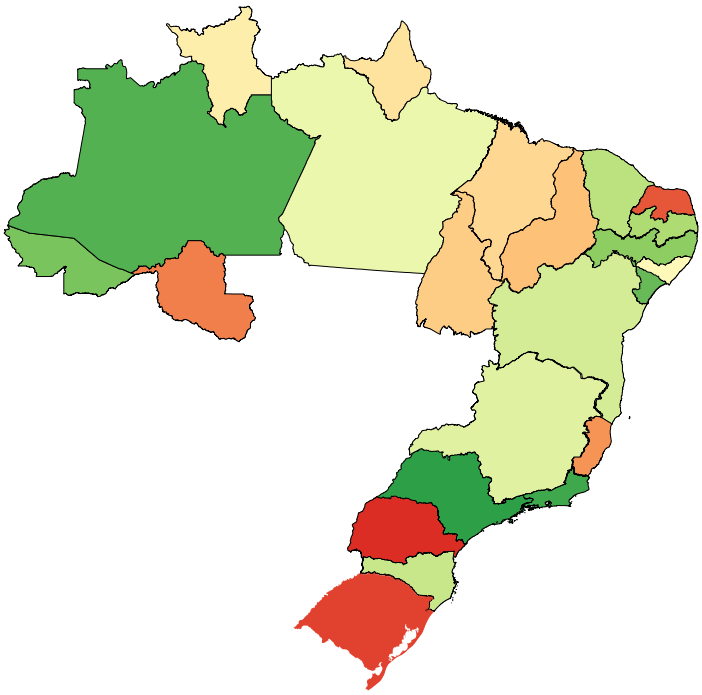

Fig 11. ECI map of the Brazilian states. The colors in the map vary from green (high ECI) to red (low ECI) and they show the variation of the ECI across the Brazilian states.

https://doi.org/10.1371/journal.pone.0197616.g011

GDP_p in the considered period. This observation can be important for the evaluation of perspectives of economic growth for Brazilian states, and, more generally, for developing countries.

The time evolution of the ranking obtained through the Exogenous Fitness algorithm shows that developed states in the top part of the ranking change little their positions, with a smooth slow motion. On the contrary states in the inferior part of the ranking changes drastically their position during the analyzed time-interval. These facts are probably due to the stability of the developed states that are in the high predictability region of the Fitness-GDP_p plane and the instability of the states in the low predictability region.

Finally, we showed the non-correlation ($\tilde{\rho}_{ExECI} = -0.14$) between the ranking obtained though the Exogenous Fitness algorithm and the results of the ECI published by Dataviva [27]. Analyzing qualitatively the ranking of the states according to ECI, we argued that this ranking appears quite unrealistic. Therefore, we propose here the Exogenous Fitness algorithm as its valid substitute. Instead, comparing the Exogenous and (Endogenous) Fitness we obtained a strong correlation ($\tilde{\rho}_{ExEn} = 0.97$) for what concerns the ranking of states. This result shows that the two algorithmic tools are almost similar in identifying the ranking of the states, but

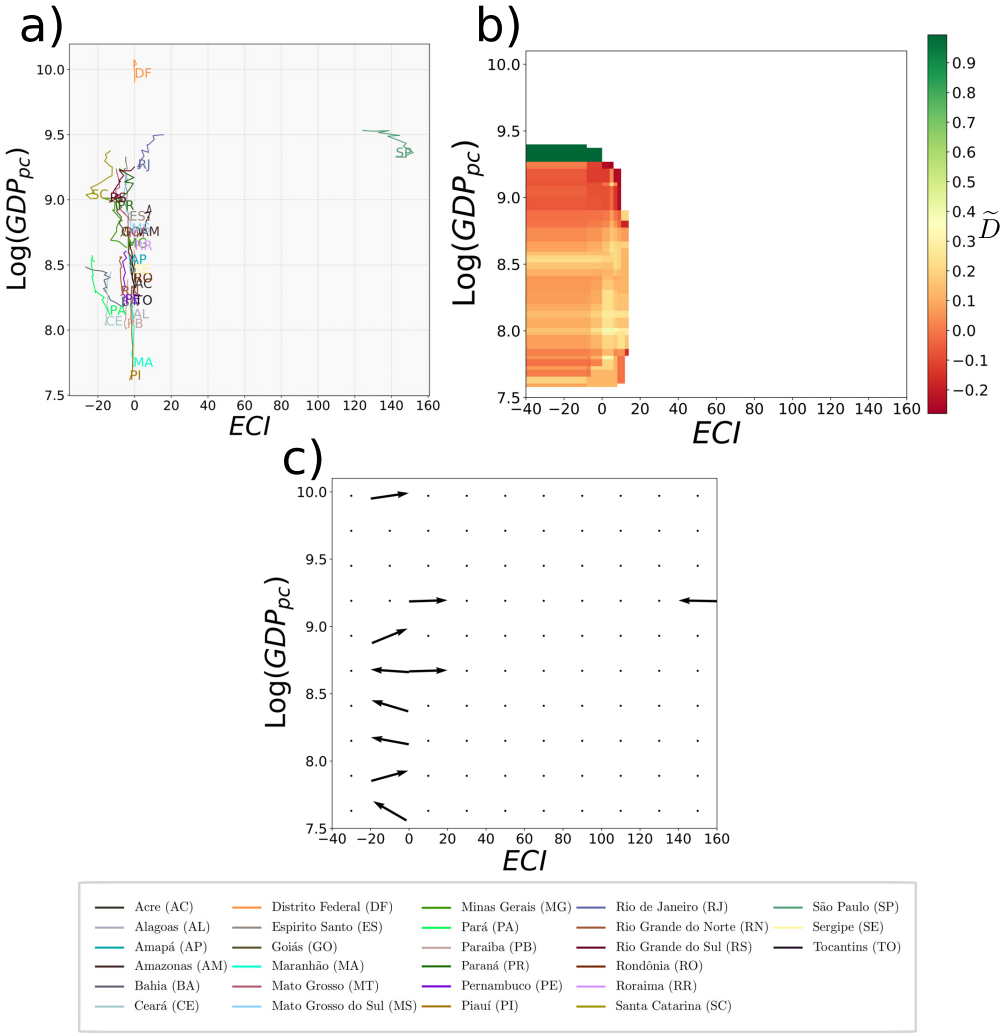

Fig 12. Evolution of Brazilian states in the ECI-Income plane. *a*) The figure shows the dynamics (from 2002 to 2015) of the Brazilian states in the ECI-Income plane, where the GDP_p is in logarithmic scale. Only the state of São Paulo and the Distrito Federal appear to be clearly distinguishable from the rest of the states. All the others states are indeed concentrated in a small region of the graph. *b*) The figure shows the coefficient $\widetilde{D}$ calculated considering the time interval 2003-2013. Colors vary from green (where the versors tend to be parallel), to red (where the versors tend to be unevenly directed). From the figure we can therefore verify that there is a low predictability of the evolution of all the states. *c*) Here we show a grid where for each cell we calculate the versor of the sum vector. From the figure we see that there is no privileged direction, indeed the vectors are unevenly directed.

https://doi.org/10.1371/journal.pone.0197616.g012

just the Exogenous Fitness algorithm provides also stable quantitative values of the Fitness, in addition to the ranking.

Appendix A

Dataset

The vast majority of data used in this paper is published by *DataViva* [27]. It is an open access platform that easily allows the access to a large amount of Brazilian socioeconomic data. The database is provided by the Brazilian *Ministries*: of *Employment* (MTPS), *Development*, *Industry and International Trade (MDIC)* and *Education* (MEC). The project is an initiative of the *Government of the state of Minas Gerais*, *Minas Gerais Investment*, *Trade Promotion Agency* (INDI) and the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG) [27, 38] in collaboration with the *Sistema Mineiro de Inovaçao* (SIMI) [39], *Big Data Corp* [40] and the *MIT Media Lab* [41]. The first version was published in November 2013 and the last one, the 3.0 version, in May 2015.

The platform includes data about imports/exports products, trade partners, occupation, economic activities, basic education, higher education and universities. All data are available in several levels of aggregation: region, state, mesoregion, microregion and municipality. The crossover among data and level of aggregations allows users to access more than 1 billion visualizations.

The visualization is made through some graph types, such as: Tree Map, Stacked, Geo Map, Network, Rings, Scatter, Compare, Occugrid, Line, Box Plot and Bar Chart. Furthermore, each data and aggregations is downloadable, and easily accessible through the API architecture [42].

Here, we use the export data of each Brazilian state for the entire time interval from 2000 to 2015. Furthermore, *DataViva* provides the data of total GDP and the total population for each state for the same time interval. Combining these with the GDP deflator GDP_{defl}, published by the *World Bank* [43], we find the real GDP *per capita* of each state as:

$$GDP_p^{real} = \frac{1}{N}\frac{GDP}{GDP_{defl}}100, \tag{7}$$

where N is the total population of each state.

Concerning World export data, used to define the matrix M_{cp} of the World countries and to calculate the products complexity, we use data from BACI dataset [29] that is grounded on the COMTRADE dataset [30]. The database, in its extension, contains data about more than 200 countries and 5000 products classified according to a 4 digit code with categorization *Harmonized System* 2007 [44]. Data are extracted from the year 2000 to 2015. The time evolution of the GDP *per capita* of each country is published by *World Bank* [45].

Acknowledgments

This work was funded by Crisis Lab (a research project financed by Italian Government) and by the Brazilian agencies: CNPq, CAPES, and FUNCAP. We especially thank our colleagues and friends L. Napolitano, A. Tacchella, A. Zaccaria, A. Patelli, G. Cimini, L. Ortenzi, and A. Sbardella for the help and the discussions.

Author Contributions

Conceptualization: Felipe G. Operti, Emanuele Pugliese, José S. Andrade, Jr., Luciano Pietronero, Andrea Gabrielli.

Data curation: Felipe G. Operti.

Formal analysis: Emanuele Pugliese, Andrea Gabrielli.

Investigation: Felipe G. Operti, Emanuele Pugliese, Andrea Gabrielli.

Methodology: Felipe G. Operti, Emanuele Pugliese, Andrea Gabrielli.

Software: Felipe G. Operti.

Supervision: José S. Andrade, Jr., Luciano Pietronero, Andrea Gabrielli.

Visualization: Emanuele Pugliese.

Writing – original draft: Felipe G. Operti, Emanuele Pugliese, Andrea Gabrielli.

References

1. Bureau of Economic Analysis. Regional Data. U.S. Department of Commerce. Retrieved 6th June 2017. Available from: https://www.bea.gov/.
2. Sawe B E. Indian States By GDP. WorldAtlas. Retrieved 5th November 2017. Available from: http://www.worldatlas.com/articles/indian-states-by-gdp.html.
3. Chepkemoi J. The Richest And Poorest States Of Brazil. WorldAtlas. Retrieved 5th November 2017. Available from: http://www.worldatlas.com/articles/the-richest-and-poorest-states-of-brazil.html.
4. Hidalgo C A and Hausmann R. The building blocks of economic complexity. PNAS. 2009. 106 no 26: 10570–10575. https://doi.org/10.1073/pnas.0900943106 PMID: 19549871
5. Tacchella A, Cristelli M, Caldarelli G, Gabrielli A, and Pietronero L. A New Metrics for Countries' Fitness and Products' Complexity. Scientific Reports. 2012. 2: 723. https://doi.org/10.1038/srep00723 PMID: 23056915
6. Cristelli M, Gabrielli A, Tacchella A, Caldarelli G, and Pietronero L. Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products. PLoS ONE. 2013. 8(8): e70726. https://doi.org/10.1371/journal.pone.0070726 PMID: 23940633
7. Cristelli M, Tacchella A, and Pietronero L. The Heterogeneous Dynamics of Economic Complexity. PLoS ONE. 2015; 10(2): e0117174. https://doi.org/10.1371/journal.pone.0117174 PMID: 25671312
8. Cristelli M, Ascagne C, Tacchella A, Cader M Z, Roster K I, and Pietronero L. On the predictability of growth. Policy Research working paper. World Bank Group, Washington, D C. 2017; WPS 8117. Available from: http://documents.worldbank.org/curated/en/632611498503242103/On-the-predictability-of-growth.
9. GDP ranking. The World Bank. Retrieved 23th June 2017. Available from: http://data.Worldbank.org/data-catalog/GDP-ranking-table.
10. Brazil Population. World Meters. Retrieved 23th June 2017. Available from: http://www.worldometers.info/world-population/brazil-population/.
11. Area Territorial Brasileira. Instituto Brasileiro de Geografia e Estatística-IBGE. Retrieved 23th June 2017. Available from: http://www.ibge.gov.br/home/geociencias/cartografia/default_territ_area.shtm.
12. Largest Countries in the World (by land area). World Meters. Retrieved 23th June 2017. Available from: http://www.worldometers.info/geography/largest-countries-in-the-world/.
13. Kuznets S. Economic Growth and Income Inequality. The American Economic Review. 1955. 45, 1: 1–28.
14. Furtado C. Formação Econômica Do Brasil. Rio de Janeiro: Fundo de Cultura. 1959.
15. Baer W, Kerstenetzky I, and Villela A V. The changing role of the State in the Brazilian economy. World Development. 1973. 1, 11: 23–44. https://doi.org/10.1016/0305-750X(73)90253-2
16. Sunkel O and Girvan C. Transnational Capitalism and National Disintegration in Latin America. Social and Economic Studies. 1973. 22: 1 132–76.
17. Hartmann D, Jara-Figueroa C, Guevara M, Simoes A, and Hidalgo C A. The structural constraints of income inequality in Latin America. arXiv. 2017.
18. Sen A. *Development as Freedom.* New York: ALFRED A. KNOPF INC. 1999.
19. Cajueiro D and Tabak BM. Ranking efficiency for emerging markets. Chaos, Solitons and Fractals. 2004. 22 349–352. https://doi.org/10.1016/j.chaos.2004.02.005

20. Chang E J, Lima E J A, and Tabak B M. Testing for predictability in emerging equity markets. Emerging Markets Review. 2004. 5 295–316. https://doi.org/10.1016/j.ememar.2004.03.005
21. Hidalgo C A, Klinger B, Barabási A-L, and Hausmann R. The product space conditions the development of nations. Science. 2007. 482–487. https://doi.org/10.1126/science.1144581 PMID: 17656717
22. Reynolds C, Agrawal M, Lee I, Zhan C, Li J, Taylor P, et al. A sub-national economic complexity analysis of Australia's states and territories. Regional Studies. 2017. 1–12.
23. Chávez J C, Mosqueda M, and Gómez-Zaldívar M. Economic Complexity and Regional Growth Performance: Evidence from the Mexican Economy. The Review of Regional Studies. 2017. 47, 2: 201–219.
24. Sbardella A, Pugliese E, and Pietronero L. Economic development and wage inequality: A complex system analysis. PLoS One. 2017. 19; 12 (9):e0182774. https://doi.org/10.1371/journal.pone.0182774 PMID: 28926577
25. Chotikapanich D, Griffiths W E, and Prasada Rao D S. Estimating and Combining National Income Distributions Using Limited Data. Journal of Business & Economic Statistics. 2007. 25, 1, 97–109. https://doi.org/10.1198/073500106000000224
26. O'Neill J. Building Better Global Economic BRICs. Goldman Sachs. Global Economics Paper. 2001. n 6.
27. DataViva. Retrieved 23th June 2017. Available from: http://www.dataviva.info/en/.
28. Balassa B. Trade Liberalisation and "Revealed" Comparative Advantage. The Manchester School. 1965. 33, 2: 99–123. https://doi.org/10.1111/j.1467-9957.1965.tb00050.x
29. Gaulier G and Zignago S. Baci: International trade database at the product-level. Retrieved 23th June 2017. Available from: http://www.cepii.fr/anglaisgraph/workpap/pdf/2010/wp2010-23.pdf.
30. UN COMTRADE database. Retrieved 23th June 2017. Available from: http://comtrade.un.org.
31. Battiston F, Cristelli M, Tacchella A, and Pietronero L. How metrics for economic complexity are affected by noise. Complexity Economics. 2014; 3: 1–22.
32. Cristelli M, Tacchella A, and Pietronero L. Economic Complexity. Measuring the Intangibles. A consumer's guide. Retrieved 23th June 2017. Available from http://www.lucianopietronero.it/wp-content/uploads/2016/02/economic_complexity_flyer-_v2.1_1.pdf
33. Pugliese E, Zaccaria A, and Pietronero L, On the convergence of the Fitness-Complexity algorithm, The European Physical Journal Special Topics. 2016. 225: 1893. https://doi.org/10.1140/epjst/e2015-50118-1
34. GDP Growth (Annual %). The World Bank. Retrieved 10th October 2017. Available from: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2016&locations=BR-RU-CN-IN&start=2000.
35. Pugliese E, Chiarotti G L, Zaccaria A, and Pietronero L. Economic Complexity as a Determinant of the Industrialization of Countries: the Case of India. Retrieved 10th October 2017. Available from: http://pilhd.phys.uniroma1.it/PILgroup_Economic_Complexity/Publications_files/industrializationIndia_v4.pdf
36. Lorenz E N. Atmospheric predictability as revealed by naturally occurring analogues. Journal of the Atmospheric Sciences. 1969. 26: 636–646. https://doi.org/10.1175/1520-0469(1969)26%3C636:APARBN%3E2.0.CO;2
37. Lorenz E N. Three approaches to atmospheric predictability. Bulletin of the American Meteorological Society. 1969. 50: 345–349.
38. The Atlas of Economic Complexity. Center for International Development at Harvard University. Retrieved 23th June 2017. Available from: http://www.atlas.cid.harvard.edu.
39. Sistema Mineiro de Inovação: SIMI. Retrieved 23th June 2017. Available from: http://www.simi.org.br/.
40. Big Data Corp. Retrieved 23th June 2017. Available from: https://www.bigdatacorp.info/.
41. MIT Media Lab. Retrieved 23th June 2017. Available from: https://www.media.mit.edu/.
42. Api DataViva. Retrieved 23th June 2017. Available from: https://github.com/DataViva/dataviva-site/wiki.
43. GDP deflator. World Bank. Retrieved 23th June 2017. Available from: http://data.Worldbank.org/indicator/NY.GDP.DEFL.ZS?locations=BR.
44. Harmonized System. Retrieved 23th June 2017. Available from: http://www.wcoomd.org.
45. GDP per capita (current US dollar). World Bank. Retrieved 23th June 2017. Available from: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2016&start=1994.

REFERÊNCIAS

[1] Marr, B. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. Forbes. Mat 21, 2018.

[2] Hastie, T; Tibshirani, R.; Friedman J. Artificial Intelligence, a modern approach; The elements of statistical learning.

[3] Nadaraya, E. A. On estimating regression. Theory of Probability & Its Applications, p. 9 (1) 141 1964.

[4] Batty, M. The new science of cities. The MIT Press. Cambridge. 2013.

[5] Bettencourt, L. M. A.; Lobo, J.; Helbing, D.; Kühnert, C.; West, G. B.; Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences of the United States of America, p. 104-(17), 2007.

[6] Operti, F. G.; Oliveira, E.; Carmona, H. A.; Machado, J. C.; Andrade Jr., J.S. The light pollution as a surrogate for urban population of the US cities. Physica A, p. 0378-4371, 2017.

[7] Oliveira, E. A.; Andrade, J. S.; Makse, H. A.; Large cities are less green. Scientific reports, p. 4:4235-1, 2014.

[8] Makse, H. A.; Havlin, S.; Stanley, H. E. Modelling urban growth patterns. Nature, p. 377 (6550) 608, 1995.

[9] Rozenfeld, H. D.; Rybski, D.; Andrade, J. S.; Batty, M.; Stanley, H. E.; Makse, H. A. Laws of population growth. Proceedings of the National Academy of Sciences of the United States of America, p. 105 (48) 18702, 2008.

[10] Caminha, C.; Furtado, V.; Pequeno, T. H. C.; Ponte, C.; Melo, H. P. M.; Oliveira, E. A.; Andrade, J. S.; Human mobility in large cities as a proxy for crime. PLoS One, p. 12: 2, 2017.

[11] Rozenfeld, H. D.; Rybski, D.; Gabaix, X.; Makse, H. A. The area and population of cities: New insights from a different perspective on citie. American Economic Review, p. 101-(5): 2205, 2011.

[12] Makse, H. A.; Andrade, J. S.; Batty, M.; Havlin, S.; Stanley, H. E. Modeling urban growth patterns with correlated percolation. Physical Review E, p. 58: 7054–7062, 1998.

[13] Snell O. Die Abhängigkeit des Hirngewichts von dem Körpergewicht und den geistigen Fähigkeiten. Arch. Psychiatr., p. 23 (2): 436–446, 1892.

[14] Hidalgo, C. A.; Hausmann R. The building blocks of economic complexity. PNAS, p. 106 no 26: 10570-10575, 2009.

[15] Tacchella, A.; Cristelli, M.; Caldarelli, G.; Gabrielli, A.; Pietronero, L. A New Metrics for Countries' Fitness and Products' Complexity. Scientific Reports, p. 2: 723, 2012

[16] Cristelli, M.; Gabrielli, A.; Tacchella, A.; Caldarelli, G.; Pietronero, L. Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products. PLoS ONE, p. 8(8): e70726, 2013.

[17] Cristelli, M.; Tacchella, A.; Pietronero, L. The Heterogeneous Dynamics of Economic Complexity. PLoS ONE, p. 10(2): e0117174, 2015.

[18] Operti F. G.; Pugliese E.; Andrade J. S. Jr; Pietronero L.; Gabrielli A. Dynamics in the Fitness-Income plane: Brazilian states vs World countries. PLOS ONE, p 13(6): e0197616, 2018.

[19] Cajueiro, D.; Tabak, B. M. Ranking efficiency for emerging markets. Chaos, Solitons and Fractals, p. 22 349–352, 2004.

[20] Chang, E. J.; Lima, E. J. A.; Tabak, B. M. Testing for predictability in emerging equity markets. Emerging Markets Review. p. 5 295 – 316, 2004.

[21] Hidalgo, C. A.; Klinger, B.; Barabási, A. L.; Hausmann, R. The product space conditions the development of nations. Science, p. 482-487, 2007.

[22] Reynolds, C.; Agrawal, M.; Lee, I.; Zhan, C.; Li, J.; Taylor, P.; Mares, T.; Morison, J.; Angelakis, N.; Roos, G. A sub-national economic complexity analysis of Australia's states and territories. Regional Studies, p. 1-12, 2017.

[23] Chávez, J. C.; Mosqueda, M.; Gómez-Zaldívar, M. Economic Complexity and Regional Growth Performance: Evidence from the Mexican Economy. The Review of Regional Studies, p. 47, 2: 201-219, 2017.

[24] Sbardella, A.; Pugliese, E.; Pietronero, L. Economic development and wage inequality: A complex system analysis. PLoS One, p. 19; 12 (9), 2017.

[25] Chotikapanich, D.; Griffiths, W. E.; Prasada, Rao D. S. Estimating and Combining National Income Distributions Using Limited Data. Journal of Business & Economic Statistics, p. 25, 1, 97-109, 2007.

[26] Falchi, F.; Cinzano, P.; Duriscoe, D.; Kyba, C. C. M.; Elvidge, C. D.; Baugh, K.; Portnov, B. A.; Rybnikova, N. A.; Furgoni, R.; The new world atlas of artificial night sky brightness, Science Advances, p. 2 (6), 2016.

[27] Kerenyi, N. A.; Pandula, E.; Feuer, G. Why the incidence of cancer is increasing: The role of light pollution. Medical Hypotheses, p. 33 (2), 1990.

[28] Blask, D. E.; Brainard, G. C.; Dauchy, R. T.; Hanifin, J. P.; Davidson, L. K.; Krause, J. A.; Sauer, L. A.; Rivera-Bermudez, M. A.; Dubocovich, M. L.; Jasser, S. A.; Lynch, D. T.; Rollag, M. D.; Zalatan, F. Melatonin-depleted blood from premenopausal women exposed to light at night stimulates growth of human breast cancer xenografts in nude rats. Cancer Research, p. 65 (23), 2005.

[29] Reiter, R. J.; Gultekin, F.; Manchester, L. C.; Tan, D. Light pollution, melatonin suppression and cancer growth. Journal of Pineal Research, p. 40 (4), 2006.

[30] Navara, K. J.; Nelson, R. J. The dark side of light at night: Physiological, epidemiological, and ecological consequences. Journal of Pineal Research, p. 43 (3), 2007.

[31] Reiter, R. J.; Tan, D.; Korkmaz, A.; Erren, T. C.; Piekarski, C.; Tamura, H.; Manchester, L. C. Light at Night, Chronodisruption, Melatonin Suppression, and Cancer Risk: A Review. Critical Reviews in Oncogenesis, p. 13 (4), 2007.

[32] Chepesiuk, R. Missing the Dark: Health Effects of Light Pollution. Environmental Health Perspectives, p. 117 (1), 2009.

[33] Salgado-Delgado, R.; Tapia Osorio, A.; Saderi, N.; Escobar, C. Disruption of circadian rhythms: A crucial factor in the etiology of depression. Depression Research and Treatment, p. 2011 (839743), 2011.

[34] Aubé, M.; Roby, J.; Kocifaj, M. Evaluating Potential Spectral Impacts of Various Artificial Lights on Melatonin Suppression, Photosynthesis, and Star Visibility. PLoS ONE, p. 8 (7), 2013.

[35] United Nations, http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html, World's population increasingly urban with more than half living in urban areas, accessed: 2017-06-01, 2014.

[36] Bettencourt, L. M. A., West, G. B. A unified theory of urban living. Nature, p. 467 (7318), 2010.

[37] Bettencourt, L. M. A.; Lobo, J.; Strumsky, D.; West, G. B. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. PLoS ONE, p. 5 (11), 2010.

[38] Melo, H. P. M.; Moreira, A. A.; Batista, É.; Makse, H. A.; Andrade, J. S. Statistical signs of social influence on suicides, Scientific reports, p. 4 (6239), 2014.

[39] Bettencourt, L. M. A.; Lobo, J. Urban scaling in Europe. Journal of The Royal Society Interface, p. 13 (116), 2016.

[40] US Census Bureau, http://www.census.gov, accessed: 2017-06-01, 2014.

[41] Doxsey-Whitfield, E.; MacManus, K.; Adamo, S. B.; Pistolesi, L.; Squires, J.; Borkovska, O.; Baptista, S. R. Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4, Papers in Applied Geography, p. 1 (3), 2015.

[42] Socioeconomic data and application center (SEDAC), http://sedac.ciesin.columbia.edu, accessed: 2017-06-01, 2016.

[43] Mills, S.; Weiss, S.; Liang, C. VIIRS day/night band (DNB) stray light characterization and correction. SPIE Optical Engineering Applications, p. 8866 (88661P), 2013.

[44] National Centers for Environmental Information (NCEI), http://www.ngdc.noaa.gov/eog/viirs.html, Visible Infrared Imaging Radiometer Suite (VIIRS), accessed: 2017-06-01, 2016.

[45] National Aeronautics and Space Administration (NASA), http://npp.gsfc.nasa.gov/viirs.html, Visible Infrared Imaging Radiometer Suite (VIIRS), accessed: 2017-06-01, 2016.

[46] Shimrat, M. Algorithm 112: Position of point relative to polygon. Communications of the ACM, p. 5 (8), 1962.

[47] Montgomery, D. C.; Peck, E. A.; Vining, G. G. Introduction to linear regression analysis. John Wiley & Sons, 2015.

[48] Watson, G. S. Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A, p. 359, 1964.

[49] Li, X.; Wang, X.; Zhang, J.; Wu, L. Allometric scaling, size distribution and pattern formation of natural cities, Palgrave Communications, p. 1 (15017), 2015.

[50] Bureau of Economic Analysis. Regional Data. U.S. Department of Commerce. Retrieved 6th June 2017. Available from: https://www.bea.gov/.

[51] Sawe, B. E. Indian States By GDP. WorldAtlas. Retrieved 5th November 2017. Available from: http://www.worldatlas.com/articles/indian-states-by-gdp.html.

[52] Chepkemoi, J. The Richest And Poorest States Of Brazil. WorldAtlas. Retrieved 5th November 2017. Available from: http://www.worldatlas.com/articles/the-richest-and-poorest-states-of-brazil.html.

[53] Cristelli, M.; Ascagne, C.; Tacchella, A.; Cader, M. Z.; Roster, K. I.; Pietronero, L. On the predictability of growth. Policy Research working paper. World Bank Group, Washington, D C. WPS 8117. Available from: http://documents.worldbank.org/curated/en/632611498503242103/On-the-predictability-of-growth, 2017.

[54] GDP ranking. The World Bank. Retrieved 23th June 2017. Available from: http://data.Worldbank.org/data-catalog/GDP-ranking-table.

[55] Area Territorial Brasileira. Instituto Brasileiro de Geografia e Estatística-IBGE. Retrieved 23th June 2017. Available from: http://www.ibge.gov.br/.

[56] Largest Countries in the World (by land area). World Meters. Retrieved 23th June 2017. Available from: http://www.worldometers.info/geography/largest-countries-in-the-world/.

[57] Kuznets, S. Economic Growth and Income Inequality. The American Economic Review, p. 45 1: 1-28, 1955.

[58] Furtado C. Formação Econômica Do Brasil. Rio de Janeiro: Fundo de Cultura, 1959.

[59] Baer, W.; Kerstenetzky, I.; Villela, A. V. The changing role of the State in the Brazilian economy. World Development, p. 1 11: 23-44, 1973.

[60] Sunkel, O.; Girvan, C. Transnational Capitalism and National Disintegration in Latin America. Social and Economic Studies, p. 22: 1 132-76, 1973.

[61] Hartmann D.; Jara-Figueroa, C.; Guevara, M.; Simoes, A.; Hidalgo, C. A. The structural constraints of income inequality in Latin America. arXiv, 2017.

[62] Sen A. *Development as Freedom*. New York: ALFRED A. KNOPF INC, 1999.

[63] O'Neill, J. Building Better Global Economic BRICs. Goldman Sachs, Global Economics Paper, n. 6, 2001.

[64] DataViva. Retrieved 23th June 2017. Available from: http://www.dataviva.info/en/.

[65] Balassa, B. Trade Liberalisation and "Revealed" Comparative Advantage. The Manchester School. p. 33, 2: 99-123, 1965.

[66] Gaulier, G.; Zignago, S. Baci: International trade database at the product-level. Retrieved 23th June 2017. Available from: http://www.cepii.fr/anglaisgraph/workpap/pdf/2010/wp2010-23.pdf.

[67] UN COMTRADE database. Retrieved 23th June 2017. Available from: http://comtrade.un.org.

[68] Battiston, F.; Cristelli, M.; Tacchella, A.; Pietronero, L. How metrics for economic complexity are affected by noise. Complexity Economics, p. 3: 1-22, 2014.

[69] Cristelli, M.; Tacchella, A.; Pietronero, L. Economic Complexity. Measuring the Intangibles. A consumer's guide. Retrieved 23th June 2017. Available from http://www.lucianopietronero.it/

[70] Pugliese, E.; Zaccaria, A.; Pietronero, L. On the convergence of the Fitness-Complexity algorithm. The European Physical Journal Special Topics, p. 225: 1893, 2016.

[71] GDP Growth (Annual %). The World Bank. Retrieved 10th October 2017. Available from: https://data.worldbank.org/indicator/

[72] Pugliese E, Chiarotti G L, Zaccaria A, and Pietronero L. Economic Complexity as a Determinant of the Industrialization of Countries: the Case of India. Retrieved 10th October 2017. Available from: http://pilhd.phys.uniroma1.it/PILgroup_Economic_Complexity/Publications_files/industrializa

[73] Lorenz, E. N. Atmospheric predictability as revealed by naturally occurring analogues. Journal of the Atmospheric Sciences, p. 26: 636-646, 1969.

[74] Lorenz, E. N. Three approaches to atmospheric predictability. Bulletin of the American Meteorological Society, p. 50: 345-349, 1969.

[75] The Atlas of Economic Complexity. Center for International Development at Harvard University. Retrieved 23th June 2017. Available from: http://www.atlas.cid.harvard.edu.

[76] Sistema Mineiro de Inovação: SIMI. Retrieved 23th June 2017. Available from: http://www.simi.org.br/.

[77] Big Data Corp. Retrieved 23th June 2017. Available from: https://www.bigdatacorp.info/.

[78] MIT Media Lab. Retrieved 23th June 2017. Available from: https://www.media.mit.edu/.

[79] Api DataViva. Retrieved 23th June 2017. Available from: https://github.com/DataViva/dataviva-site/wiki.

[80] GDP deflator. World Bank. Retrieved 23th June 2017. Available from: http://data.Worldbank.org/indicator/

[81] Harmonized System. Retrieved 23th June 2017. Available from: http://www.wcoomd.org.

[82] GDP per capita (current US dollar). World Bank. Retrieved 23th June 2017. Available from: https://data.worldbank.org/indicator/.

[83] Chodrow, P. S. Structure and information in spatial segregation. Proceedings of the National Academy of Sciences, p. 114 (44): 11591-11597, 2017.

[84] Lichter, D. T.; Parisi, D.; Taquino, M. C. Toward a New Macro-Segregation? Decomposing Segregation within and between Metropolitan Cities and Suburbs. American Sociological Review, p. 80 (4): 843-873, 2015.

[85] Fowler, C. S. Segregation as a multiscalar phenomenon and its implications for neighborhood-scale research: the case of South Seattle 1990–2010. Urban Geography, p. 37 (1): 1-25, 2016.

[86] Boustan L P. Racial Residential Segregation in American Cities. The Oxford Handbook of Urban Economics and Planning, 2012.

[87] Readon, S. F.; Farrell, C. R.; Matthews, S. A.; O'Sullivan, D.; Bischoff, K.; Firebaugh, G. Race and space in the 1990s: Changes in the geographic scale of racial residential segregation, 1990–2000. Social Science Research, p. 38: 55-70, 2008.

[88] Readon, S. F.; Mathhews, S. A.; O'Sullivan, D., Lee, B.; Firebaugh, G.; Farrell, C. R.; Bischoff, K. The geographic scale of metropolitan racial segregation. Demography, p. 45 (3): 489-514, 2008.

[89] Charles, C. Z. The dynamics of Racial Residential Segregation. Annual Review of Sociology. p. 29: 167-207, 2003.

[90] Massey, D. S.; Denton, N. A. American Apartheid. Harvard Univ Pr. isbn-10: 0674018214.

[91] Roberto, E.; Hwang, J. Barriers to Integration: Physical Boundaries and the Spatial Structure of Residential Segregation. arXiv. 2017.

[92] Firebaugh, G.; Farell, C. R. Still Large, but Narrowing: The Sizable Decline in Racial Neighborhood Inequality in Metropolitan America, 1980–2010. Demography, p. 53 (1): 139-64, 2016.

[93] Logan, J. R.; Stults, B. J. The Persistence of Segregation in the Metropolis: New Findings from the 2010 Census. Accessed 9th of March 2018. url: https://s4.ad.brown.edu/Projects/Diversity/Data/Report/report2.pdf, 2010.

[94] Logan, J. R.; Stults, B. J.; Farley, R. Segregation of minorities in the metropolis: Two decades of change. Demography. p. 41 (1): 1-22, 2004.

[95] Roberto, E. The Divergence Index: A Decomposable Measure of Segregation and Inequality. arXiv, 2015.

[96] Louf, R.; Barthelemy, M. Patterns of residential segregation. PLoS ONE. p. 11 (6), 2016.

[97] Roberto, E. The Spatial Proximity and Connectivity (SPC) Method for Measuring and Analyzing Residential Segregation. arXiv, 2016.

[98] Jargowsky, P. A.; Kim, J. A Measure of Spatial Segregation: The Generalized Neighborhood Sorting Index. National Poverty Center Working Paper Series. Accessed the 12th March 2018. url: http://www.npc.umich.edu/publications/working_papers/.

[99] Readon, S. F.; O'Sullivan, D. Measures of Spatial Segregation. Sociological Methodology. p. 34: 121-162, 2004.

[100] Readon, S. F.; Firebaugh, G. Measures of Multigroup Segregation. Sociological Methodology. p. 32 (1): 33-67, 2002.

[101] Massey, D.; Denton, N. The Dimensions of Residential Segregation. Social Forces. p. 67 (2): 281-315, 1988.

[102] Winship, C. A Revaluation of Indexes of Residential Segregation. Social Forces, p. 55 (4): 1058-1066, 1977.

[103] Duncan, O. D.; Duncan, B. A Methodological Analysis of Segregation Indexes. American Sociology Rev., p. 20: 210-217, 1955.

[104] Watson, T. Inequality and the measurement of of residential segregation by income in American neighborhoods. The review of income and wealth, p. 55 (3): 820-844, 2009.

[105] Acevedo-Garcia, D.; Lochner, K. A.; Osypuk, T. L.; Subramanian, S. V. Future Directions in Residential Segregation and Health Research: A Multilevel Approach. The American Journal of Public Health, p. 93 (2): 215-221, 2003.

[106] Williams, D. R.; Collins, C. Racial Residential Segregation: A Fundamental Cause of Racial Disparities in Health. Public Health Reports, p. 116 (5): 404-416, 2001.

[107] Fang, J.; Madhavan, S.; Bosworth, W.; Alderman, M. H. Residential segregation and mortality in New York City, p. 47 (4): 469-76, 1998.

[108] Helbrecht, I. Gentrification and Displacement. SpringerVS, 2018.

[109] Brown-Saracino, J. Explicating Divided Approaches to Gentrification and Growing Income Inequality. Annual Review of Sociology, p. 43: 515-39, 2017.

[110] Lees, L. Gentrification, Race, and Ethnicity: Towards a Global Research Agenda? City and Community, p. 208-214, 2016.

[111] Freeman, L. Neighbourhood Diversity, Metropolitan Segregation and Gentrification: What Are the Links in the US? UrbanStudies, p. 46 (10): 2079-2101, 2008.

[112] Freeman, L.; Braconi, F. Gentrification and Displacement New York City in the 1990s. Journal of the American Planning Association, p. 39-52, 2007.

[113] Smith, N. The Encyclopedia of Housing. London: Taylor & Francis, 1998.

[114] Curran, W. From the Frying Pan to the Oven: Gentrification and the Experience of Industrial Displacement in Williamsburg, Brooklyn, p. 44 (8): 1427-1440, 2007.

[115] Zuk, M.; Bierbaum, A. H.; Chapple, K.; Gorska, K.; Loukaitou-Siders, A. Gentrification, Displacement, and the Role of Public Investment. Journal of Planning Literature, p. 33 (1): 31-44, 2017.

[116] Zukin, S. New Retail Capital and Neighborhood Change: Boutiques and Gentrification in New York City. City and Community, 2009.

[117] Lees, L. Super-gentrification: The Case of Brooklyn Heights, New York City. Urban Studies, p. 40 (13): 2487-2509, 2003.

[118] Schaffer, R.; Smith, N. The Gentrification of Harlem? Annals of the Associarron of Amrrirun Grogrupherr, p. 76 (3): 347-365, 1986.

[119] Vijaymeena, M. K.; Kavitha, K. A survey on similarity measures in text mining. Machine Learning and Applications: An International Journal (MLAIJ), p. 3 (1), 2016.

[120] Gini C. Variabilità e mutabilità. Reprinted in Pizetti. Salvemini. Memorie di metodologica statistica. Rome: Libreria Eredi Virgilio Veschi, 1955.

[121] Manson, S.; Schroeder, J.; Riper, D. V.; Ruggles, S. IPUMS National Historical Geographic Information System: Version 12.0 [Database]. Minneapolis: University of Minnesota, 2017.

[122] US Census Bureau. Accessed 9th of March 2018. url: ttp://www.census.gov.