SCIENTIFIC REPORTS
nature research

OPEN

Reconstructing commuters network using machine learning and urban indicators

Received: 17 May 2019
Accepted: 1 August 2019
Published online: 13 August 2019

Gabriel Spadon[1], Andre C. P. L. F. de Carvalho[1], Jose F. Rodrigues-Jr[1] & Luiz G. A. Alves[1,2]

Human mobility has a significant impact on several layers of society, from infrastructural planning and economics to the spread of diseases and crime. Representing the system as a complex network, in which nodes are assigned to regions (*e.g.*, a city) and links indicate the flow of people between two of them, physics-inspired models have been proposed to quantify the number of people migrating from one city to the other. Despite the advances made by these models, our ability to predict the number of commuters and reconstruct mobility networks remains limited. Here, we propose an alternative approach using machine learning and 22 urban indicators to predict the flow of people and reconstruct the intercity commuters network. Our results reveal that predictions based on machine learning algorithms and urban indicators can reconstruct the commuters network with 90.4% of accuracy and describe 77.6% of the variance observed in the flow of people between cities. We also identify essential features to recover the network structure and the urban indicators mostly related to commuting patterns. As previously reported, distance plays a significant role in commuting, but other indicators, such as Gross Domestic Product (GDP) and unemployment rate, are also driven-forces for people to commute. We believe that our results shed new lights on the modeling of migration and reinforce the role of urban indicators on commuting patterns. Also, because link-prediction and network reconstruction are still open challenges in network science, our results have implications in other areas, like economics, social sciences, and biology, where node attributes can give us information about the existence of links connecting entities in the network.

Humans move daily to work, do business, have leisure, meet people, and perform routine activities. Modeling human mobility is vital to better allocate resources and to improve the impacts of human activities in the community (nearby people) and the environment (cities and nature). From physics and mathematics to geography and social sciences, several researchers have tried different approaches to understand the impacts of human movement on society and vice-versa[1,2]. Human mobility is shaped by the urban organization[3] and can also change cities as an effect of traffic congestion[4]. Mobility patterns are associated with energy use[5,6], the spread of diseases[7–10], the occurrence of crimes[11,12], and others. Therefore, good predictive models can, for instance, help improve daily human activities with better urban planning, and also help policy-makers with more informed decisions to intervene in the disease spreading and crime.

Prediction of migration from one area to another, at different time scales, is one of the most critical challenges in human mobility. It is common to represent the system as a spatial complex network, where each node represents a region (*e.g.*, city) and the links between two regions indicate the flow of people. Thus, physics-inspired models such as the gravitation[13,14] and radiation models[15,16] have been used to predict the edges of the network as well as their weights (the number of people migrating). These models assume that the number of people going from one region/node to another decays with the distance separating them and is proportional to the populations' masses of these regions[13–16]. However, this assumption often fails to accurately describe the flow of people because of other factors that can increase or decrease mobility, such as the underlying transportation network[17], socio-economic aspects[3,18–20], and transport congestion[4]. Usually, these models are limited to predicting the weights of existing links, and they overestimate the number of connections between nodes when dealing with sparse

[1]University of Sao Paulo, Institute of Mathematics and Computer Sciences, Sao Carlos, SP, 13566-590, Brazil. [2]Northwestern University, Department of Chemical and Biological Engineering, Evanston, IL, 60208-3112, USA. Correspondence and requests for materials should be addressed to G.S. (email: spadon@usp.br) or L.G.A.A. (email: lgaalves@northwestern.edu)

mobility networks. Therefore, these limitations make it hard to generalize the model on unseen data and to reconstruct the structure of human mobility networks.

Link prediction and network reconstruction is a very active area of research in network science[21]. Most of the literature on link prediction is based on evaluating the similarity between nodes and suggesting missing links. For instance, the metrics used to evaluate the existence of links include, but are not limited to, the number of common neighbors[22,23], the existence of short paths between nodes[24], measures of centrality[25], and hierarchical structures within the network[26]. Reconstruction techniques also include methods based on maximum-entropy distribution regarding the node degree and the flow strength as constraints[27,28] and Bayesian inference of links based on edge data information[29,30]. In general terms, these methods use network-based metrics to infer the existence (or non-existence) of links, which significantly differs from predictive models based on meta-data attributes such as the population size or the distance between nodes.

A few models can be found in the literature where the node's attributes are used as complementary information to predict links. For instance, in the context of social contact networks, mobility patterns such as co-location of individuals was used to evaluate the probability of individuals to be connected in a social network[31], restaurant reviews were used to predict links in a taste similarity network[32], and the frequency of programming code commits was used to predict the mobility in cyberspace of projects[33]. The focus of our work is in the class of methods that can use node's attributes, such as urban metrics, as input data to fit models that can predict links and can be further extended to other data sets where the dependent variable is unknown.

Recently, an unprecedented amount of data related to human behavior and cities became available, from GPS tracking of mobile phones to census data of thousands of people[14,34–37]. These data promoted intense research on human mobility[1], including transportation networks[38], commuters networks[17], and network models of migration[39]. On the other hand, urban indicators were used to describe scaling in cities[3], to measure the performance of cities[40] and the similarity among different municipalities[41,42], and to describe crime-related phenomena[3,11,43]. However, a connection between human mobility and urban indicators, such as unemployment rate and Gross Domestic Product (GDP) is still missing. Understanding the influence of these indicators on the individual choices on daily commuting to work could help us to predict the flow of people between different areas and reconstruct the structure of the commuters network. Such a gap has led us to frame the question we want to answer: *how to quantify the number of people commuting in between cities taking into account a more comprehensive set of indicators, beyond distance and population, that might attract or repel commuters?*

In this study, we propose an alternative approach to deal with the reconstruction of commuters networks using supervised machine learning to understand the relationships between urban indicators and the structure of commuters networks. We perform an analysis based on 22 urban indicators of 5,565 Brazilian municipalities, together with the daily number of people commuting between every city in the data set. We show that the gravitation and radiation models have limited predictive power to classify whether there is a link between two cities or not and to quantify the flow (number of people commuting) between cities. In contrast, we show that predictions based on machine learning algorithms and urban indicators can reconstruct the commuters network with 90.4% of accuracy and describe 77.6% of the variance observed in the flow of people between cities. Further, we interpret the machine learning results using SHapley Additive exPlanations (SHAP) values[44] to quantify the importance of the urban indicators in predicting human mobility. We show that distance is a critical metric for predicting human mobility, but other indicators, like GDP and unemployment rate, also play a significant role in attracting/repelling people to a particular area. Our approach provides a better way to quantify human mobility and shed new lights on the role of urban indicators on commuting patterns. Because link-prediction and network reconstruction are still open challenges in network science, our results have implications in other areas, like economics, social sciences, and biology, where node attributes can give us information about the existence of links connecting network's entities.

Results

We started our study from the observation that people move from one city s to work in another city t, taking into consideration the advantages of living nearby or far away from where they work. In Fig. 1A each node represents a city that offers different opportunities (in terms of jobs), costs of living (number of houses available), and other factors, such as hospitals, schools, and urban parks, to name a few. Two given cities, s and t, are separated by a distance r_{st}, and people have a cost to commute from their home city to the city where they work. Existing migration models assume that the flow of people from city s to city t is proportional to their populations and decays with the distance. However, plenty of other factors play an important role when deciding where to live or work.

To quantify the number of people commuting in between cities taking into account a more comprehensive set of indicators, beyond distance and population, we collected data about the number of people commuting from one city to another (pendular migration), in 2010, considering all the 5,565 Brazilian cities, together with 22 urban indicators related to these municipalities. This data is provided and maintained by the Brazilian Institute of Geography and Statistics (IBGE)[45]. The data consists of the number of people that are living in a city s and commuting to work in a city t daily, and of urban indicators describing the cities in terms of population, labor, tuition rate, economy, sanitation, infrastructure, and violence, to name a few. We suppose that these indicators can be associated with a high or low number of commuters. Thus, less economically developed cities, usually those with fewer job opportunities, inferior infrastructure, and possibly higher indices of violence, would attract a lower number of people. On the other hand, those with a higher income, larger population, better social development, and infrastructure, might offer better job opportunities and would attract a higher number of people.

To investigate the flow of people commuting from one city to another, we assigned a node for each city and, for each non-zero flow, an edge with weight w_{st} equal to the number of people commuting from city s to t in the year 2010. Figure 1B illustrates the Brazilian commuters network using a kernel-based edge bundling technique[46]. To illustrate our urban indicators, in Fig. 1C, we present the Pearson correlation coefficient between the 22 urban indicators.

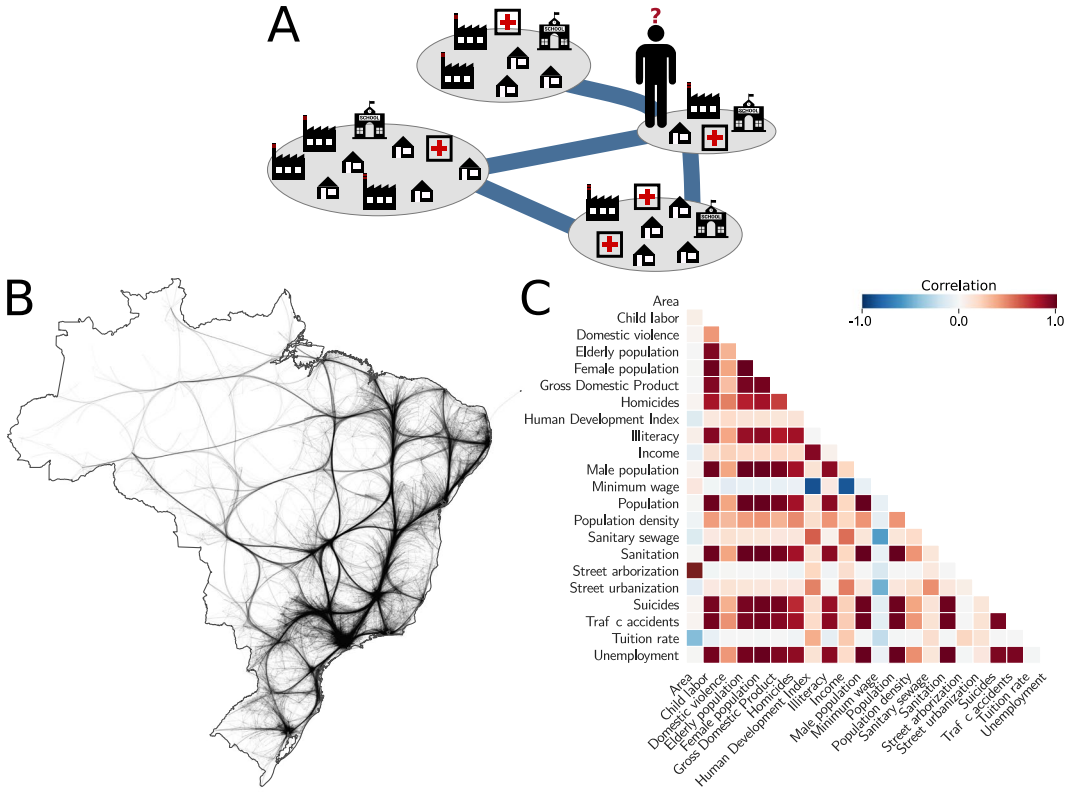

Figure 1. Commuters network and urban indicators related to human mobility. (**A**) Illustration of the work-mobility problem when people have to choose where to work based on distance, job opportunities, housing prices, and other variables related to urban systems. (**B**) Brazilian commuters network illustrated through a kernel-based Edge Bundling technique[46] where the thickness of the edges represents the flow intensity. (**C**) Correlation analysis of Brazilian urban indicators considering those that describe the population, labor, tuition rate, economy, sanitation, infrastructure, violence, and others. Reddish colors correspond to positively correlated indicators and bluish colors to negatively correlated ones.

State-of-the-art models. Next, we investigate the limitations of two models widely used to describe the flow of people between cities, namely, the gravitation model and the radiation model.

Gravitation model. Inspired by Newton's law of universal gravitation, it has been first used to predict the trading between nations; the model is defined as:

$$T_{st}^{GM} = C\frac{m_s^{\alpha} n_t^{\beta}}{r_{st}^{\gamma}}, \tag{1}$$

where m_s and n_t are the supply and demand forces of nations s and t, respectively[47]. This model was widely applied to predict population movement[14], economic relations between countries[48], cargo shipping volume[49], and long-distance phone calls[50]. In the context of migration, T_{st}^{GM} is the number of people commuting, s and t are cities, m_s and n_t are the population masses of these cities, and r_{st} is the distance between them. Thus, the interaction between two municipalities is directly proportional to the population masses and inversely proportional to the distance. Such a model depends on the parameters C, α, β, and γ, which can be estimated through Ordinary Least Squares (OLS) applying the logarithm operation on both sides of the formula. Although the gravitation model provides a good fit for a wide variety of scenarios, there are still some unsolved problems. For instance, Simini *et al.*[15] pointed out that:

- The model's equation lacks a rigorous mathematical derivation;
- The augmentation of the model is unrestricted, what would scale the number of parameters;
- The model fails when the data is not sufficient to estimate its parameters;
- The flow is only a function of population and distance, not including any other characteristic inherent to the cities and their neighbors;
- The number of commuters increases without limit as the population of the target city grows, becoming higher than the total population of the source city; and,
- The model is deterministic, being unable to estimate fluctuations in the number of people commuting.

Radiation model. Formulated in terms of the radiation and absorption processes from physics, this model was proposed to solve the problems inherent to the gravitation model[15]. It is based on a diffusion process in which an object in a specific location emits particles, having nearby objects with a different probability of absorbing these particles, which varies according to the forces acting upon them. In the context of commuters, the objects are the cities, and the particles are the people commuting. Thus, cities are radiating commuters, which are absorbed by neighboring cities.

In contrast to the gravitation model, in the radiation model, the distance between cities is not the major limitation to commuters; differently, the limitation arises from the supply and demand of commuters concerning the neighboring cities. The idea of supply and demand comes from the theory of intervening opportunities; this theory defines that mobility patterns are more influenced by the commuters' opportunity to establish in the target city than the distance to such a city[19]. The approximate number of intervening opportunities p_{st} is calculated according to the population of the source s and target t cities. It also takes into account the population of all the neighbors of the source city in a maximum radius of up to the distance to the target city. Thus, the radiation model can be defined as:

$$T_{st}^{Rad} = T_s \frac{m_s n_t}{(m_s + p_{st})(m_s + n_t + p_{st})}, \tag{2}$$

where m_s is the population of the source city, n_t the population of the target city, and p_{st} is the total population in the circle of radius r_{st} centered at city s.

The only parameter of the radiation model, T_s, is the scaling factor of the outgoing number of commuters from the source city. The value of T_s can be estimated by adjusting the equation $T_s = \alpha m_s$ to the data using OLS, in which m_s is the population of the source city and α is the intersection point of the fitting between the population size and the number of commuters from every city in the whole data set. The major limitation of the model is that T_s depends on the number of commuters. Thus, the model requires the correct information (or an estimation) about the number of people commuting between cities, which makes it difficult to generalize the model to unseen data.

Similarly, Ren *et al.*[17] proposed the cost-based radiation model to predict the flow of commuters in spatial networks. Although the model has been reported as more efficient than its predecessor, its predictive ability is still limited to cases where there is information about the existence of links. As a consequence, the radiation model and its cost-based version are not able to reconstruct the structure of the commuters network.

It is also worth to mention that most of these models (gravitation and radiation models) were validated through correlation analysis, rather than using the variance between the predicted and real values; this fact causes an overestimation of the capability of the model in predicting the flow of people between cities. The variance is more error-sensitive than the correlation coefficient (*e.g.*, Pearson, Spearman, or Kendall). Moreover, when assessing the results, some authors take into account only existing edges to compare with the predictions, causing, once more, an optimistic estimation of their predictive performance.

Next, to further confirm our claims, we applied the gravitation and radiation models to predict the number of people in the Brazilian commuters network. Our first observation is that these models (Eqs 1 and 2) are not able to reconstruct the unweighted projection of the commuters network structure, because any pair of cities with non-zero population would have an edge, resulting in a fully connected network. If one considers faraway cities, it is possible to have flow close to zero, but for typical distances between Brazilian cities (average value of $\sim$1,000 km), we would still find a non-zero flow. The resulting networks generated by these models are far from the sparse networks observed in real data. Thus, we quantified the number of commuters considering only edges that we know beforehand to have a non-zero flow. We considered the distance to be the length of the geodesic path (in kilometers) between two given cities, which is calculated from their coordinates provided by the Brazilian census data[45].

Our evaluation started by fitting the models (Eqs 1 and 2) to the data through OLS to estimate their descriptive parameters. Next, we evaluated the results using R^2-score (coefficient of determination) and Pearson correlation coefficient. The R^2-score measures the variance of a dependent variable that was predicted using independent variables; the metric is defined in the range $]-\infty,\ 1]$. An R^2-score closer to 1, means that the variance of the prediction concerning the real value is low[51]. Notice that this metric can be negative since the model can be arbitrarily wrong. It is worth mentioning that, along with the text, we express the R^2-score through percentage (*e.g.*, 14% instead of 0.14) because we use it to discuss the proportion of variance explained by the correlation between the predicted and observed flow. The Pearson correlation coefficient ρ, in turn, measures the linearity of two variables regarding each other. The metric is defined in the range $[-1,\ 1]$, in which -1 indicates a perfect negative correlation, 0 indicates no linear correlation and 1 indicates a perfect positive correlation[52]. With very wrong predictions, one could find results that are very correlated with the empirical values ($\rho \approx 1$), misleading the evaluation of the predictive model performance.

Figure 2 presents the results related to the fitting of the gravitation and radiation models to our data. From this figure, we verify that the values predicted by both models are positively correlated (values greater than 0.60) with the real data, but the R^2-score is meaningless regardless of the model (values below 0.25). Despite the good correlation, the models were not able to correctly predict the data, and the values differ sharply from the real ones. The same conclusion was brought by Masucci *et al.*[16], who analyzed the ability of generalization and universality of both models with a data set of commuters in the surroundings of London, UK. In agreement with what we observed when evaluating the model using the Brazilian data set, neither model could satisfactorily fit their data. Thus, we confirmed that the gravitation and radiation models could not correctly describe the mobility pattern of the intercity commuters network.

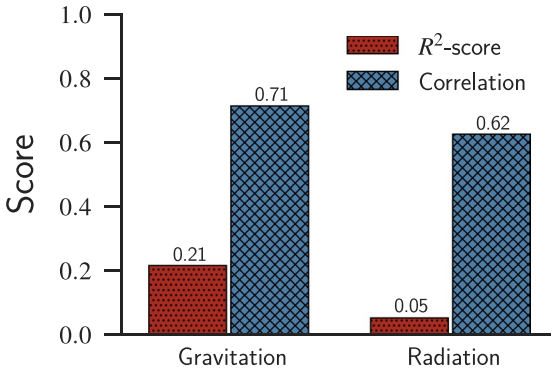

Figure 2. Evaluating the predictive ability of the gravitation and radiation models. We used two metrics, the coefficient of determination (referred to as R^2-score), and the Pearson correlation coefficient; both metrics consider the predicted flow and the observed flow of the models. Although the results show a linear correlation between the predicted and observed flows, the R^2-score reveals that the predicted values are still very noisy when compared to the real ones, suggesting that neither model is accurate enough in reconstructing the commuters network.

Alternative modeling using machine learning. Next, we propose the use of predictive models induced by machine learning algorithms to predict the number of intercity commuters and to reconstruct the structure of the corresponding commuters network. Our approach differs from the regular link prediction because it does not take into account any network-based topology metric as an independent variable. Instead, we reconstruct the network based on the distance r between two cities, their populations' size and a set of urban indicators related to each city. To do so, we investigate two predictive tasks:

- Classification: induce a binary classifier able to predict whether a link between a pair of cities exists or not;
- Regression: induce a regressor able to predict the number of commuters between a pair of cities.

Reconstructing the commuters network structure using machine-learning classification. In our context, a binary classifier predicts whether a link between a city s and a city t exists or not based on the distance r_{st}, populations sizes m_s and n_t, and more 21 urban indicators from each city, $U_i = \{u_{i_0}, \ldots, u_{i_{20}}\}$, $i \in \{s, t\}$, and $u_{i_j} \in \mathbb{R}$ for $0 \leq j \leq 20$, resulting in a total of 45 urban indicators (referred to as features). That is, given an ordered pair $\langle city_s, city_t \rangle$ whose urban indicators define a set $S_{st} = \{r_{st}, m_s, n_t, U_s, U_t\}$, $S_{st} \in \mathbb{R}^{|S|}$, a binary classifier refers to a function

$$Class: \mathbb{R}^{|S|} \rightarrow \{0, 1\} \tag{3}$$

in which, 0 indicates no link between s and t, and 1 indicates the opposite.

To find the best classifier, we first employed the holdout approach, splitting the data into 70% for training and 30% for testing. Then, we sampled the training data using stratified k-fold cross-validation, with $k = 5$. The training data was used in the model selection, feature selection, and hyperparameter tuning. The test set was used only in the final step, after the hyperparameter tuning, to evaluate the predictive performance, generalization, and universality of the model in an unseen data set.

Subsequently, we tested 34 classification algorithms with default hyperparameter values from the scikit-learn[53] and eXtreme Gradient Boosting (XGBoost)[54] libraries, from which only 27 were able to fit the data and provide the resulting accuracy. The 7 removed algorithms (see Supplementary Table 1) fail to fit the data (*i.e.*, to converge) without a pre-tuning of hyperparameters, which is intentionally not covered by our methodology. We do not perform the pre-tuning of hyperparameter because it would exponentially increase the processing time of the experiments' pipeline with no improvement guaranteed.

Following, to select the best among the 27 remaining classifiers, we used bootstrapping sampling to evaluate their predictive performance following an accuracy-based perspective. This procedure consists in feeding the model with randomly selected samples to assess its variance. Figure 3A shows the results of the classifiers that passed our first test. The algorithms were sorted in ascending order according to their predictive accuracy. It is worth to mention that the best algorithms (rightmost) are the CatBoost, XGBoost, and Light Gradient Boosting Machine (LGBM), and the ones with the worst performance are the Perceptron, Passive Aggressive Classifier, and Gaussian Naive Bayes (NB). The average accuracy score in the experiment varies from 50.2% in the worst case to 87.9% in the best case — see Supplementary Table 3. Notice that, the algorithm with the highest median score, lowest variance, and fewer outliers is the CatBoost. However, the difference between the CatBoost (first-placed) and XGBoost (second-placed) is irrelevant (see Supplementary Table 3), and the XGBoost is around fifty times faster per train iteration than the CatBoost. For this reason, we have chosen the XGBoost as the best algorithm for this problem.

Following, we evaluated the learning curve considering only the best classifier (*i.e.*, XGBoost), which showed an average accuracy score of 87.5% on the training set, with little variance and no significant accuracy outliers. The learning curve assesses the model predictability by varying the size of the training set. The curve in Fig. 3B

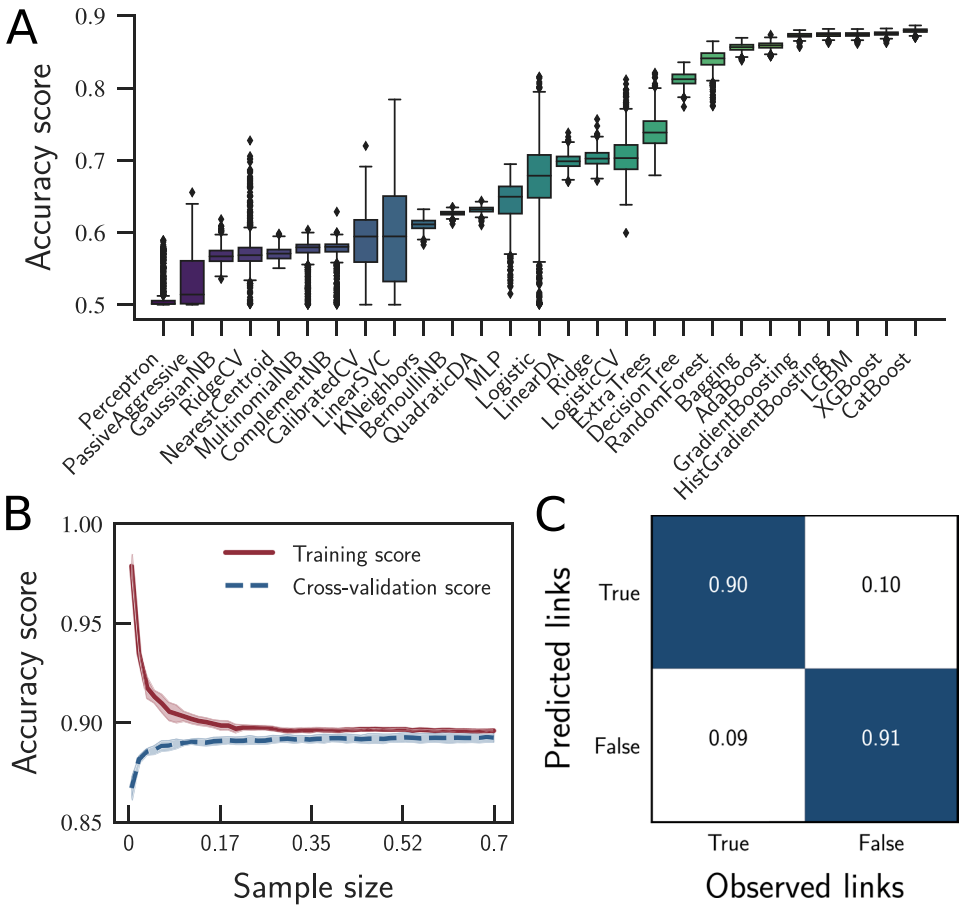

Figure 3. Performance of the classification algorithms in reconstructing the unweighted projection of the commuters network; see Supplementary Table 1 for a list of the classifiers' acronyms. (**A**) The accuracy of the classification algorithms in determining whether a link between a given pair of cities exists or not. (**B**) XGBoost learning curves varying the sample size. (**C**) XGBoost confusion matrix after hyperparameter tuning, showing an overall accuracy of 90.4%. Specifically, the results showed 90% of true positives, 10% of false negatives, 9% of false positives, and 91% of true negatives.

shows that 89% of the training set is the right amount of data required to train the classifier and a classifier trained up to such amount of data yields an accuracy of about 89.2%.

To reduce the computational cost in the induction of the predictive model using XGBoost, we used a threshold-based feature selection (see Methods), measuring the accuracy of the model by progressively removing features. We considered the degree of importance of a feature as the number of times the feature is used to split a node into two trees in the XGBoost algorithm. By doing so, we were able to remove 24 predictive features without reducing the model's predictive accuracy (see Supplementary Table 5).

Finally, we used the test set (the remaining 30% of the data previously not used) to test the induced model. We first calculated the predictive accuracy of our classifier on the test data using the model with no grid-search optimization, finding 89.3% of accuracy. The grid-search was able to improve this value by 1.1% (see Methods), which, in a scenario with thousands of cities with millions of possibilities of commuters flowing between them, means more hundreds of thousands of correctly predicted links. Therefore, such an improvement resulted in an overall accuracy of 90.4%, as shown in the confusion matrix in Fig. 3C.

Reconstructing the weighted commuters network using machine-learning regression. Deeper in the analytical scenario, we are interested not only in knowing whether a link exists or not but also in the weight w_{st} of each link, which represents the flow of people from city s to t. That is, given an edge $\langle city_s, city_t \rangle$ and its corresponding set of indicators $S_{st} = \{r_{st}, m_s, n_t, U_s, U_t\}, S_{st} \in \mathbb{R}^{|S|}$, we seek a function

$$Weigh: \mathbb{R}^{|S|} \rightarrow \mathbb{N} \tag{4}$$

where *Weight* predicts the number of commuters between cities s and t.

Similarly to the prediction of links, we first used holdout to split the data into 70% for training and 30% for testing. Subsequently, the training set was used for 5-fold cross-validation, model selection, feature selection, and hyperparameter tuning, and the test set to evaluate our model in an unseen data set.

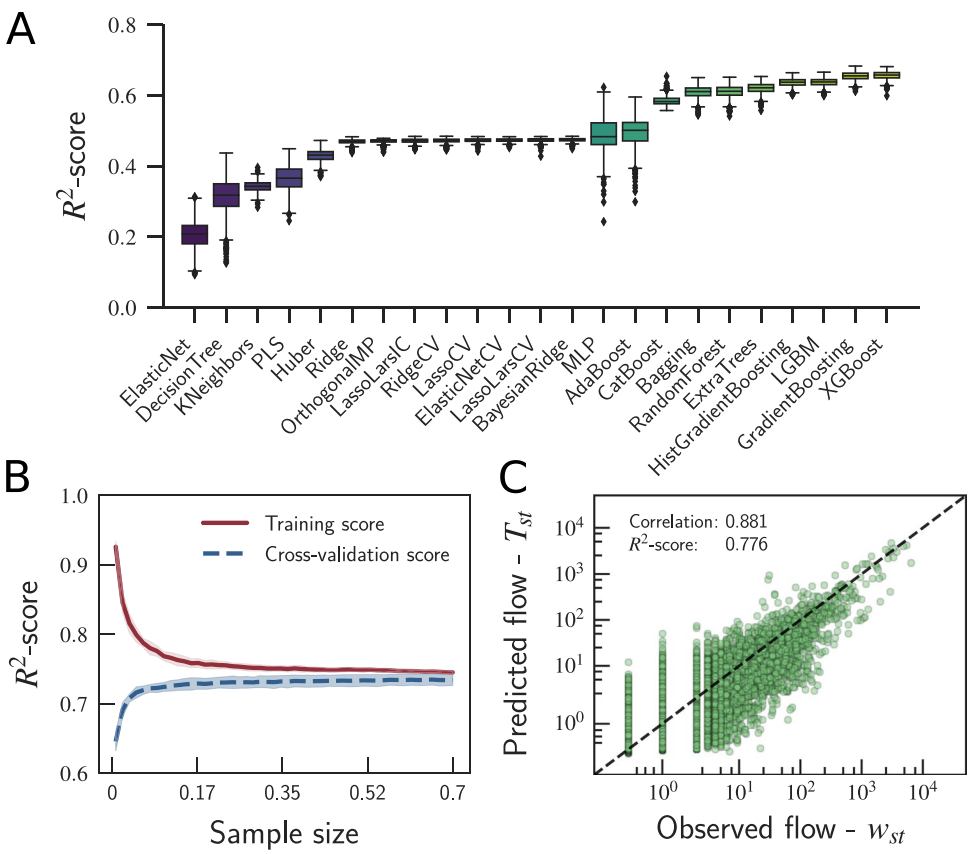

Figure 4. Performance of the regression algorithms in reconstructing the weighted projection of the commuters network; see Supplementary Table 2 for a list of the regressors' acronyms. (**A**) R^2-score of the regressors' performance in quantifying the number of people (*i.e.*, $Weight_{st}$) commuting from city s to city t. (**B**) XGBoost learning curves varying the sample size. (**C**) Predictions of $Weight_{st}$ made by the XGBoost regressor, after hyperparameter tuning, compared with the actual flow w_{st}. Specifically, our model as able to achieve an R^2-score of 77.6% and a Pearson correlation coefficient of 0.881.

We used bootstrapping sampling to assess the variance of different regressors in terms of the R^2-score. From the 44 models, 21 were not able to provide all values (including outliers) of R^2-score higher than 0 (see Supplementary Table 2). Figure 4A shows the results obtained from the 23 models that passed the first test, all of which were tested with default hyperparameter values from the scikit-learn[53] and XGBoost[54] libraries. The models are presented in ascending order according to the value of their R^2-score. The results reveal that the three best algorithms are XGBoost, Gradient Boosting, and LGBM, and the three worst are Elastic-Net, Decision Tree, and K-Nearest Neighbors. The average R^2-score in the test varies from 20.6% in the worst case (*i.e.*, Elastic-Net) to 65.6% in the best case (*i.e.*, XGBoost), see Supplementary Table 4.

From now on, our focus will be on the best regression model, which was induced by XGBoost with an average R^2-score of 65.6%, with no significant variance after a thousand predictions using different random samples of the training set. The further reason to choose the XGBoost as the best algorithm is that it is fast, as it provides a parallel tree boosting that solves problems faster than its competitors and has support to solve problems on high-performance computing environments. Thus, we evaluated the learning curve of the XGBoost regressor (see Fig. 4B), assessing the model predictability by varying the size of the training set, and we found that 100% of the training data is required to train the regressor, yielding an R^2-score of 73.4%. One can see that the regressors' predictions are more challenging than those of the binary classification algorithms, requiring more data during the training phase and yielding more errors.

Finally, we used the 30% data reserved for testing the model. The results indicated that the regressor was able to predict 73.1% of the variance observed in the data. After the hyperparameter tuning (see Methods), the XGBoost regressor described 77.6% of the variance. Although the gains seem to be small, our model is intended to be used on data sets of thousands of cities, which answer for millions of possibilities of commuters flowing between pairs of cities. In such a case, 4.5% improvement means weights closer to the real ones among the actual links. Figure 5C compares the predictions about the number of commuters ($Weight_{st}$) with the actual values (w_{st}) observed in the data. Notice that, our model can recover the weighted network structure and predict the number of commuters with much higher precision than the gravitation and radiation models. In our case, even considering links with $w_{st} = 0$, our predictions are much more accurate than those from the gravitation and radiation models (see Fig. 2), which were induced using only existing flows.

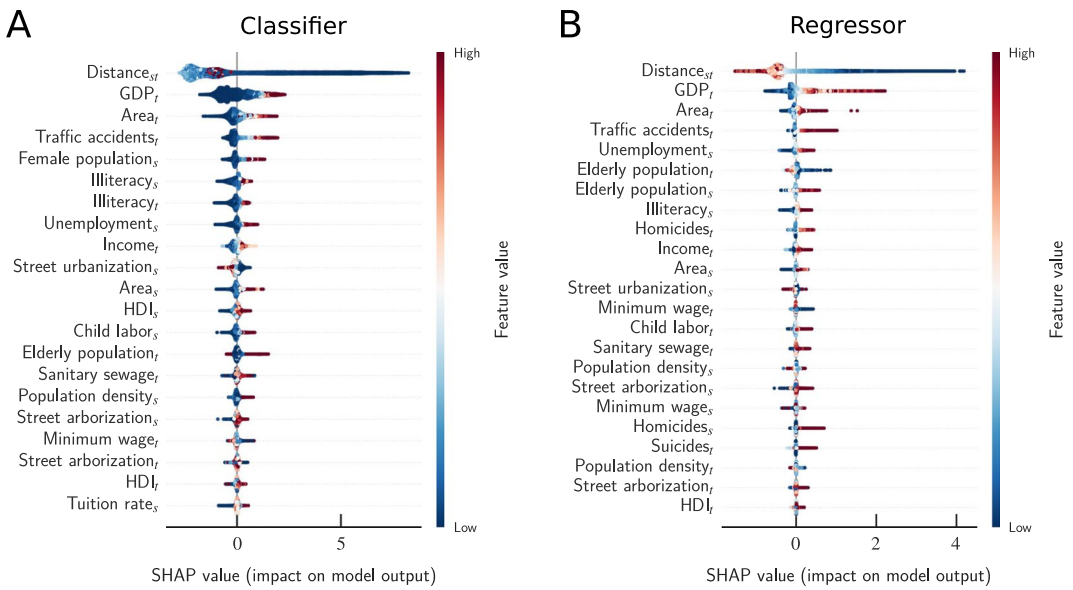

Figure 5. Interpreting the relationship between flow and features. Analysis of feature importance in the XGBoost (**A**) classifier and (**B**) regressor using SHAP values. The features are ranked by importance in descending order based on the sum of the SHAP values over all the samples. The violin-shaped plots show the distribution of the impacts of each feature on every point of the model output. The colors represent the SHAP value, varying from low values (blueish colors) to high values (reddish colors).

Interpreting machine learning models using SHapley additive exPlanations (SHAP). In contrast with the gravitation and radiation models, the proposed models can accurately reproduce the structure of the network, predicting whether a link between a given pair of cities exists or not (XGBoost classification) and accurately estimate the number of commuters between two cities (XGBoost regressor), reconstructing the weighted structure of the commuters network. The price of having a more complex model is that it becomes harder to interpret the relationship between the independent variables and our predictor. However, if we could understand how the algorithm makes decisions, the decision process would be a valuable resource to learn about the equations that describe our systems. Thus, instead of assuming mathematical relationships and trying to fit the model to the data, we could look into the results and try to figure out what the data tell us about the model that describes our system and the relationships between the commuter's flow and our features.

To turn our black-box algorithm in a more interpretative model, we calculated the features' importance and the impact of the features on individual predictions using the SHapley Additive exPlanations (SHAP) values[44,55,56]. The SHAP metric is based on the Shapley values introduced by Lloyd Shapley in 1953 in the context of game theory[57]; it helps one to understand how the model decides to make a prediction and which features contribute to improving the accuracy of the model. The course of action of SHAP is to calculate the importance of a feature by comparing what the model predicts with and without the feature. Notice that the order in which we add new features to the model can affect its predictions. Thus, we have to permute over all the possible feature orderings to fully capture the impact of a feature on the model.

In Fig. 5A, we show the features considered important for the XGBoost classifier, as well as the distribution of the impacts of each feature on the model output. While high values of distance $Distance_{st}$ (also known as r_{st}) are not good predictive features for the presence of links (values smaller than 0), high GDP_t values increase the predictive accuracy of the model (values larger than 0). High rates of the elderly population in the target city also strongly affect the model accuracy. Finally, the total urban area or traffic accidents are more important than the population density.

In Fig. 5B, we show the features' importance for the XGBoost regressor, as well as the distribution of the impacts of each feature on the model output. In this case, high values of distance $Distance_{st}$ have a low impact on the predictive performance of the model (values smaller than 0), whereas high GDP_t increases the predictive performance of the model (values larger than 0). Higher values of the elderly population in the target city decrease the performance of the algorithm. Again, we verify that several urban indicators have a role in defining the flow of people from one city to others and that urban indicators, such like GDP, area, and traffic accidents, are more useful to improve the predictive performance of the model than population size.

Discussion

Predicting decisions related to individual choices, such as housing and working places, is a difficult task. These decisions are tied to many variables that are often based on personal reasons and cannot be easily measured. However, the analysis of SHAP values can help us to understand how urban indicators and distance can influence this decision process and what makes people commute from one area to another to work. The SHAP values point

four features that are mutually important in both classification and regression models, *i.e.*, Distance, GDP, Area, Traffic accidents in the target city, followed by other less important features.

Distance is the most important feature in predicting the existence of a link between a pair of cities and quantifying the number of people commuting from one city to the other. It is traditionally included in migration models. Distant cities make the journey to work unfeasible, as workers must return home by the end of the day, limiting the commuting to a certain distance. However, because of other factors, cities close to each other may have an insignificant flow, either because they offer similar opportunities or because the infrastructure and transport limit the commuting.

GDP (Gross Domestic Product) is the second most prominent feature. The cities with the highest GDP have the highest demand for commuters and the best job opportunities. As a consequence, they attract more commuters. Looking at Fig. 5, it is possible to observe that the higher the GDP value, the higher its contribution to the predictive performance of the XGBoost induced models. Another possible explanation is that these cities are more productive because they attract more workers and, therefore, they have higher GDP values, as suggested by Keuschnigg *et al.*[58] for long-term migration.

Area is related to the physical size of a city, and it is the third most influential variable. Larger cities, like metropolis and megalopolis, usually are hubs of industries and enterprises and, consequently, tend to offer more job opportunities. This variable fails to affect the flow prediction when cities have a wide territorial extension but are used for other purposes (*e.g.*, natural reserves, forest, and huge urban parks) rather than urban expansion. Our data set reveals just a few records of cities occupying a large area with no significant number of commuters. These are the cases of cities with SHAP values close to 0, as we can observe in Fig. 5B.

Traffic accident is the fourth most important variable. This indicator was selected by the model because it has a high correlation with a higher flow of people and automobile vehicles (see Fig. 1C), which is also a trait related to the economics of the city, as is the case of GDP. Recall that, although the features we used are correlated, the feature selection process chose to keep the ones that it considers to be important, removing the other ones without negatively impact the predictive performance of the model (classifier and regressor).

The following features are ordered differently, depending on the predictive task, whether it is classification or regression. In the classification task, the classifier does not require much to predict a fair number of links correctly. For example, using just the previously discussed selected features on the training set, the classifier presented an accuracy close to 87.6%. The other features improved the model accuracy by 1.6%, also on the training set. Among all variables, the *unemployment* indicator of the source city (*i.e.*, where people live), seems to have a significant impact on the existence of flow between two cities. High rates of unemployment seem to make commuters leave the city where they live (*i.e.*, repel) to work in nearby cities with better job opportunities (*i.e.*, absorb).

Differently from the classification task, in the regression task, the algorithm demands more information to estimate the value of flow close to the actual value. Using only the four most important features, the model describes 70.6% of the variance of the data on the training set. The remaining features are responsible for improving the regressor's predictive performance by 2.8% on the training set. The *elderly population* indicator has an interesting behavior: the larger the number of elderly people in the home city and the smaller their number in the city of work, the higher the flow of commuters between the two.

The remaining urban indicators showed little relevance in predicting the flow of people, despite having a significant impact on the predictions of the classification and regression tasks. This set of features together provides a better prediction and highlights the importance of taking into account a more comprehensive set of indicators, in addition to the typical metrics used in migration models, such as population and distance. Further, such models could be enhanced by adding other indicators that were not explored in our data set, which could potentially improve the predictions of links and flow of commuters.

Finally, we believe that our approach is not only to be used on the task of commuters network reconstruction but also to complex networks of different domains where node's attributes can give us information about the existence of links connecting them. For instance, this methodology could be applied in economic trade networks[59,60], where the amount of money exchanged could be predicted in terms of indicators as GDP, unemployment, interest rate, production, corporate profits, and other macroeconomic variables; in social networks[61], where friendship connections could be unveiled by individual node's preferences, such as music or sport preferences, language or individual socioeconomic status and education; in metabolic networks[62], where biochemical reactions (links) could be predicted in terms of metabolites chemical, physical, and biological properties. Thus, further insights from these networks could be obtained by exploring the SHAP values to identify decisive indicators to reconstruct each network topology.

Methods

Data set. We collected data about the number of people commuting from a home city to another city to work (pendular migration) in the year of 2010, considering all the 5,565 Brazilian cities as well as 22 yearly-updated urban indicators. This data is provided and maintained by the Brazilian Institute of Geography and Statistics (IBGE)[45]. The data was modeled as a complex network represented as a directed graph $G = \{V, E\}$ composed of 5,565 vertices and 55,247 edges with non-zero weights. A vertex $s \in V$ is considered to be a city, and an edge $e \in E$ is an ordered pair $e = \langle s, t \rangle$ representing the flow from the source city $s \in V$ to the target city $t \in V$. The existence of commuters from city s to city t is noted as $Class_{st} = 1$, and $Class_{st} = 0$ indicates the opposite; the flow $Weight_{st}$ is an integer representing the number of commuters that move daily from city s to city t. The urban indicators shape an initial 45-value feature vector. This vector is composed of urban indicators (a single value per indicator) from the source and target cities (22 values for each city) and their distance (in kilometers).

Data split. The data set composed by the number of commuters (target feature) and the 2×22 urban indicators plus the distance between pairs of cities (predictive features) was split in a 70/30 ratio preserving the ratios of existing and non-existing links in both sets. Thus, 70% was used as a training set, and the other 30% was used as a test set to evaluate the predictive performance, generalization, and universality of the model induced after all training and selection steps. To assure that the model could learn all inherent nuances of our data (*e.g.*, different intercity road systems), we also stratified the data by the Federal Brazilian States to feed the model with a proportional amount of information about every state. This also saves computational time to train the models by eliminating most of the non-existent links across states.

Class balancing. Because imbalanced data sets favor the majority class, we balanced the classes of the resulting data set keeping 50% of the data labeled with existing links, $w_{st} \neq 0$, and the other with non-existing links $w_{st} = 0$, resulting in 110,494 pairs of cities (edges). For each existing connection between a city from state A and another from state B, our data set has a non-existing link between cities (chosen at random) from these states. This process was carried out for both tasks, classification, and regression. In the classification, the labels are $Class_{st} = 1$ if there is a non-zero flow and $Class_{st} = 0$ otherwise, whereas, in the regression task, the data was labeled with the number of commuters in the flow from one city to the other.

Training set stratification. We used a label-based stratified k-fold with 5 folds for the classification tasks (categorical data) and a regular k-fold with 5 folds for the regression tasks (number of commuters).

Model selection. Using the training set, which represents 70% of the whole data set and is composed of 77,345 edges, we tested 78 algorithms of machine learning, including 34 classifiers and 44 regressors. In this step, we used the validation subset of the training data to select the best classifier to predict the existence of a non-zero flow (link) between cities and the best regressor to predict the number of commuters flowing between them (links' weight). This task was carried out using Bootstrapping, which is a statistical resampling method that consists of sampling with replacement to define the upper and lower bound of a statistical evaluation metric (accuracy and R^2-score)[63]. We used this method to evaluate different models over random samples of the data set, without compromising the significance of the results. From all classifiers and regressors, we only used those that could yield a predictive performance without needing early hyperparameter tuning. Almost one-third of all algorithms did not meet such constraints, leaving 27 classifiers and 23 regressors for the experiments. The remaining algorithms went through a bootstrapping sampling by training each one with tiny random samples (1% of the data chosen with replacement, *i.e.*, 7,734 edges) of the training set (70% of the whole data set, *i.e.*, 77,345 edges) to access the variance of predictions over a thousand iterations. Through the bootstrapping sampling results, we defined the upper and lower bound of the scoring metric (accuracy and R^2-score). Using these metrics as criteria, we select the classifier and regressor with the highest median, lowest variance, and fewer outliers. It is important to note that bootstrap sampling uses random samples of 1% of the training set to increase the randomness within each test sample so to capture all nuances related to the variance and outliers among thousands of predictions.

eXtreme gradient boosting (XGBoost). XGBoost is a machine learning library that provides a set of techniques to deal with and model data through both classifiers and regressors. Such a library was designed to be highly efficient, flexible, and portable. All the model-inferring algorithms were implemented under the Gradient Boosting framework providing further support to parallel and distributed processing[54]. Gradient Boosting, on the other hand, renders a predictive model through tree ensembles, which are submitted to optimization through an arbitrary differentiable loss function[64].

Learning curves. We used analysis of learning curves to assess the prediction capability of each model (*i.e.*, XGBoost classifier, and regressor) when increasing the size of the training data in increments of 1% up to 70%. The right amount of data to train each model is the one with the best trade-off between bias and variance, that is, where the training score curve and the cross-validation curve have the lowest deviation.

Feature selection. We performed a threshold-based feature selection to decrease the complexity of the induced model by removing features from the least important to the most important one up to a dynamically-defined importance-threshold. Such a process provides a subtler interpretation of the features' impact on the model's output and is carried out together with the XGBoost algorithm, which provides an importance score to each feature during the training phase. The importance score is defined in the interval $[0, 1]$, where 0 indicates a feature with minimum or no importance, and 1 indicates the opposite. The importance score consists of counting the number of times each feature is used to split a node into two trees by the XGBoost algorithm during the training phase. The feature selection process consists of measuring the model's prediction score while progressively removing the features from the least important to the most important one. These iterations define a threshold that splits the interval into two, such that features with importance between 0 and the threshold are removed. The threshold is defined by iteratively increasing the value from 0 up to 1 until the model's prediction score using all features starts decreasing. It is important to note that the features removed are not relevant to the induced model because they do not add useful information to predict the target data. This comes from the fact that urban indicators are collinear to each other, which means each indicator might contain traces of other indicators within it. Bettencourt[3] already discussed such phenomena on urban indicators, however, there still no way to avoid or remove it completely from such type of data. The threshold-based feature selection works around this problem by removing non-relevant and collinear features, which absence impacts the model nor positively neither negatively. Therefore, the benefit of using feature selection, in this case, is its ability to reduce the model complexity by decreasing the number of dependable variables.

Hyperparameter tuning. For XGBoost classifier and regressor, we performed hyperparameter tuning; a brute-force technique that seeks for the set of hyperparameters within a pre-defined hyperparameters' subspace that optimizes a provided machine learning model. For the sake of experimentation, we tested 3 million possibilities out of a hyperparameters space that is exponentially larger. Such a fine-tuning is considerably time-consuming and can take several weeks of intensive computing for each model. On the other hand, the time is proportional to the number of tested hyperparameters, which is a fair trade-off, and that can be decreased with high-performance computing.

For our experiments, we found that the best hyperparameter values for the XGBoost classifier were: *(i)* learning rate = 0.05; *(ii)* number of trees = 300; *(iii)* maximum tree depth = 9; *(iv)* fraction of observations used as random samples by each tree = 0.85; *(v)* subsample ratio of features when constructing each tree = 0.8; and, *(vi)* regularization term of the model's weights = 0.75. In the case of the XGBoost regressor, the best hyperparameter values were: *(i)* learning rate = 0.05; *(ii)* number of trees = 900; *(iii)* maximum tree depth = 7; *(iv)* fraction of observations used as random samples by each tree = 0.85; *(v)* regularization term of the model's weights = 1.25; *(vi)* minimum loss reduction required to split a node into two trees = 0.25; and, *(vii)* minimum weight needed in a child node = 2.

Model evaluation. After all the training and selection steps, we used the testing set to compare the non-optimized and optimized models, reporting the results in a confusion matrix for the classifier and a scatter plot for the regressor. The classifiers were evaluated through the balanced accuracy score (*i.e.*, the ratio between the correct classified instances and all instances, normalized by the number of elements per label) and the regressors through the R^2-score (*i.e.*, the variance between the predicted and observed values).

SHAP values and feature importance. Given a specific prediction $f(x)$, we calculate the importance of a feature by comparing what a model predicts with and without the feature. Because the order in which we add new features to the model can affect its predictions, we permuted over all possible feature ordering. Mathematically, the Shapley value for a particular feature i (out of M total features), given a prediction x is:

$$\phi_i(f, x) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)], \tag{5}$$

where S is the set of features used to induce the model, and f_x is the prediction considering the indicated set of features[55]. In practice, this is too difficult to be calculated because there are far too many possible combinations. Instead, we used the SHAP library to calculate ϕ_i, which is optimized to take advantage of different model's structures[44]. Thus, the rank of feature importance is given by the sum of SHAP value magnitudes ϕ_i over all samples. This procedure allows a more in-depth interpretation of the impact of the features on the prediction of individual data, turning the black-box algorithm in a more interpretable model.

References

1. Barbosa, H. *et al*. Human mobility: Models and applications. *Physics Reports* **734**, 1–74, https://doi.org/10.1016/j.physrep.2018.01.001 (2018).
2. Ullman, E. L. *Geography as spatial interaction* (University of Washington Press, 1980).
3. 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* **5**, 1–9, https://doi.org/10.1371/journal.pone.0013541 (2010).
4. Louf, R. & Barthelemy, M. Modeling the polycentric pransition of cities. *Physical Review Letters* **111**, 198702, https://doi.org/10.1103/PhysRevLett.111.198702 (2013).
5. Trenchard, H. & Perc, M. Energy saving mechanisms, collective behavior and the variation range hypothesis in biological systems: A review. *Biosystems* **147**, 40–66, https://doi.org/10.1016/j.biosystems.2016.05.010 (2016).
6. Helbing, D. *et al*. Saving human lives: What complexity science and information systems can contribute. *Journal of Statistical Physics* **158**, 735–781, https://doi.org/10.1007/s10955-014-1024-9 (2015).
7. Eubank, S. *et al*. Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180, https://doi.org/10.1038/nature02541 (2004).
8. Colizza, V., Barrat, A., Barthélemy, M. & Vespignani, A. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences* **103**, 2015–2020, https://doi.org/10.1073/pnas.0510525103 (2006).
9. Balcan, D. *et al*. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* **106**, 21484–21489, https://doi.org/10.1073/pnas.0906910106 (2009).
10. Yang, H.-X., Tang, M. & Wang, Z. Suppressing epidemic spreading by risk-averse migration in dynamical networks. *Physica A: Statistical Mechanics and its Applications* **490**, 347–352, https://doi.org/10.1016/j.physa.2017.08.067 (2018).
11. Caminha, C. *et al*. Human mobility in large cities as a proxy for crime. *PLoS One* **12**, 1–13, https://doi.org/10.1371/journal.pone.0171609 (2017).
12. Spadon, G. *et al*. Behavioral characterization of criminality spread in cities. vol. 108, 2537–2541, https://doi.org/10.1016/j.procs.2017.05.118, International Conference on Computational Science, ICCS 2017, 12–14 June 2017, Zurich, Switzerland (2017).
13. Zipf, G. K. The P1P2/D hypothesis: on the intercity movement of persons. *American Sociological Review* **11**, 677–686, https://doi.org/10.2307/2087063 (1946).
14. Jung, W.-S., Wang, F. & Stanley, H. E. Gravity model in the Korean highway. *EPL (Europhysics Letters)* **81**, 48005, https://doi.org/10.1209/0295-5075/81/48005 (2008).
15. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96, https://doi.org/10.1038/nature10856 (2012).
16. Masucci, A. P., Serras, J., Johansson, A. & Batty, M. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E* **88**, 022812, https://doi.org/10.1103/PhysRevE.88.022812 (2013).
17. Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M. C. & Toroczkai, Z. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature Communications* **5**, 5347, https://doi.org/10.1038/ncomms6347 (2014).
18. Ravenstein, E. G. The laws of migration. *Journal of the Statistical Society of London* **48**, 167–235, https://doi.org/10.2307/2979181 (1885).

19. Stouffer, S. A. Intervening opportunities: A theory relating mobility and distance. *American Sociological Review* **5**, 845–867, https://doi.org/10.2307/2084520 (1940).
20. Louail, T. *et al*. Uncovering the spatial structure of mobility networks. *Nature Communications* **6**, 6007, https://doi.org/10.1038/ncomms7007 (2015).
21. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* **390**, 1150–1170, https://doi.org/10.1016/j.physa.2010.11.027 (2011).
22. Newman, M. E. Clustering and preferential attachment in growing networks. *Physical Review E* **64**, 025102, https://doi.org/10.1103/PhysRevE.64.025102 (2001).
23. Kossinets, G. Effects of missing data in social networks. *Social Networks* **28**, 247–268, https://doi.org/10.1016/j.socnet.2005.07.002 (2006).
24. Jeh, G. & Widom, J. Simrank: a measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 538–543, https://doi.org/10.1145/775047.775126 (ACM, 2002).
25. Fu, C. *et al*. Link weight prediction using supervised learning methods and its application to yelp layered network. *IEEE Transactions on Knowledge and Data Engineering* **30**, 1507–1518, https://doi.org/10.1109/TKDE.2018.2801854 (2018).
26. Clauset, A., Moore, C. & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98, https://doi.org/10.1038/nature06830 (2008).
27. Mastrandrea, R., Squartini, T., Fagiolo, G. & Garlaschelli, D. Enhanced reconstruction of weighted networks from strengths and degrees. *New Journal of Physics* **16**, 043022, https://doi.org/10.1088/1367-2630/16/4/043022 (2014).
28. Squartini, T., Mastrandrea, R. & Garlaschelli, D. Unbiased sampling of network ensembles. *New Journal of Physics* **17**, 023052, https://doi.org/10.1088/1367-2630/17/2/023052 (2015).
29. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences* **106**, 22073–22078, https://doi.org/10.1073/pnas.0908366106 (2009).
30. Peixoto, T. P. Reconstructing networks with unknown and heterogeneous errors. *Physical Review X* **8**, 041011, https://doi.org/10.1103/PhysRevX.8.041011 (2018).
31. Wang, D., Pedreschi, D., Song, C., Giannotti, F. & Barabasi, A.-L. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, 1100–1108, https://doi.org/10.1145/2020408.2020581 (ACM, New York, NY, USA, 2011).
32. Xuan, Q. *et al*. Modern food foraging patterns: Geography and cuisine choices of restaurant patrons on yelp. *IEEE Transactions on Computational Social Systems* **5**, 508–517 (2018).
33. Xuan, Q., Okano, A., Devanbu, P. & Filkov, V. Focus-shifting patterns of oss developers and their congruence with call graphs. In *Proceedings of the 22Nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE 2014, 401–412, https://doi.org/10.1145/2635868.2635914 (ACM, New York, NY, USA, 2014).
34. Makse, H. A., Havlin, S. & Stanley, H. E. Modelling urban growth patterns. *Nature* **377**, 608, https://doi.org/10.1038/377608a0 (1995).
35. Thiemann, C., Theis, F., Grady, D., Brune, R. & Brockmann, D. The structure of borders in a small world. *PLoS One* **5**, 1–7, https://doi.org/10.1371/journal.pone.0015422 (2010).
36. Roth, C., Kang, S. M., Batty, M. & Barthélemy, M. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS One* **6**, 1–8, https://doi.org/10.1371/journal.pone.0015923 (2011).
37. Barthélemy, M. Spatial networks. *Physics Reports* **499**, 1–101, https://doi.org/10.1016/j.physrep.2010.11.002 (2011).
38. Guimerà, R., Mossa, S., Turtschi, A. & Amaral, L. A. N. The worldwide air transportation network: Anomalous centrality, community structure, and cities global roles. *Proceedings of the National Academy of Sciences* **102**, 7794–7799, https://doi.org/10.1073/pnas.0407994102 (2005).
39. Lee, S. H., Ffrancon, R., Abrams, D. M., Kim, B. J. & Porter, M. A. Matchmaker, matchmaker, make me a match: Migration of populations via marriages in the past. *Physical Review X* **4**, 041009, https://doi.org/10.1103/PhysRevX.4.041009 (2014).
40. Alves, L. G. A., Mendes, R. S., Lenzi, E. K. & Ribeiro, H. V. Scale-adjusted metrics for predicting the evolution of urban indicators and quantifying the performance of cities. *PLoS One* **10**, 1–17, https://doi.org/10.1371/journal.pone.0134862 (2015).
41. Domingues, G. S., Silva, F. N., Comin, C. H. & da F Costa, L. Topological characterization of world cities. *Journal of Statistical Mechanics: Theory and Experiment* **2018**, 083212, https://doi.org/10.1088/1742-5468/aad365 (2018).
42. Spadon, G., Gimenes, G. & Rodrigues, J. F. Topological street-network characterization through feature-vector and cluster analysis. In *International Conference on Computational Science*, 274–287, https://doi.org/10.1007/978-3-319-93698-7_21 (Springer, 2018).
43. Alves, L. G. A., Ribeiro, H. V., Lenzi, E. K. & Mendes, R. S. Distance to the scaling law: A useful approach for unveiling relationships between crime and urban metrics. *PLoS One* **8**, 1–8, https://doi.org/10.1371/journal.pone.0069580 (2013).
44. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4768–4777 (Curran Associates Inc., USA, 2017).
45. Brazilian Institute of Geography and Statistics (IBGE). Accessed: 2017-09-01 (2017).
46. Moura, D. C. 3D Density Histograms for Criteria-driven Edge Bundling. *ArXiv:1504.0268* (2015).
47. Leibenstein, H. Shaping the world economy: Suggestions for an international economic policy. *The Economic Journal* **76**, 92–95, https://doi.org/10.2307/2229041 (1966).
48. Helpman, E., Melitz, M. & Rubinstein, Y. Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics* **123**, 441–487, https://doi.org/10.3386/w12927 (2008).
49. Kaluza, P., Kölzsch, A., Gastner, M. T. & Blasius, B. The complex network of global cargo ship movements. *Journal of The Royal Society Interface* **7**, 1093–1103, https://doi.org/10.1098/rsif.2009.0495 (2010).
50. Expert, P., Evans, T. S., Blondel, V. D. & Lambiotte, R. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences* **108**, 7663–7668, https://doi.org/10.1073/pnas.1018962108 (2011).
51. Carpenter, R. Principles and procedures of statistics, with special reference to the biological sciences. *The Eugenics Review* **52**, 172, https://doi.org/10.1002/bimj.19620040313 (1960).
52. Chiang, C. *Statistical Methods of Analysis*. Statistical Methods of Analysis (World Scientific, 2003).
53. Pedregosa, F. *et al*. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
54. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 785–794, https://doi.org/10.1145/2939672.2939785 (ACM, 2016).
55. Lundberg, S. M. *et al*. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* **2**, 749, https://doi.org/10.1038/s41551-018-0304-0 (2018).
56. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).
57. Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games* **2**, 307–317 (1953).
58. Keuschnigg, M., Mutgan, S. & Hedström, P. Urban scaling and the regional divide. *Science Advances* **5**, https://doi.org/10.1126/sciadv.aav0042 (2019).
59. Alves, L. G. A., Mangioni, G., Rodrigues, F., Panzarasa, P. & Moreno, Y. Unfolding the complexity of the global value chain: Strength and entropy in the single-layer, multiplex, and multi-layer international trade networks. *Entropy* **20**, 909, https://doi.org/10.3390/e20120909 (2018).

60. Alves, L. G. A. *et al*. The nested structural organization of the worldwide trade multi-layer network. *Scientific Reports* **9**, 2866, https://doi.org/10.1038/s41598-019-39340-w (2019).
61. Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Social Networks* **25**, 211–230, https://doi.org/10.1016/S0378-8733(03)00009-1 (2003).
62. Guimera, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433**, 895, https://doi.org/10.1038/nature03288 (2005).
63. Efron, B. Bootstrap methods: Another look at the jackknife. In Kotz, S. & Johnson, N. L. (eds) *Breakthroughs in Statistics: Methodology and Distribution*, 569–593 (Springer New York, New York, NY, 1992).
64. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**, 1189–1232 (2001).

Acknowledgments

The authors would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), through grants 2013/07375-0, 2014/25337-0, 2016/16987-7, 2016/17078-0, 2016/18615-0, 2017/08376-0, and 2019/04461-9; Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq), through grants 303694/2015-7, 404870/2016-3, 167967/2017-7, and 305580/2017-5; and Intel for their financial support. They also would like to thank Daniel M. Lima, Lucas C. Scabora, and Willian D. Oliveira for their help with the Edge Bundling algorithm.

Author Contributions

G.S. and L.G.A.A. conceived the experiments. G.S. conducted the experiments. G.S. and L.G.A.A. analyzed the results. G.S., A.C.C., J.F.R. and L.G.A.A. validated the results. G.S. and L.G.A.A. wrote the original draft. All authors reviewed and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-48295-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.