Machine learning for modeling mobility flows between locations

Oriol Cabanas[1], Lluís Danús[1], Esteban Moro[34], Roger Guimerà[12] and Marta Sales-Pardo[1]
[1]Dept. Chemical Engineering, Universitat Rovira i Virgili, 43007 Tarragona
[2]ICREA, 08007 Barcelona
[3]Connection Science, Institute for Data Science and Society, MIT, Cambridge, United States
[4]Department of Mathematics, Universidad Carlos III de Madrid, Spain

Modeling human mobility and understanding the main features involved have inspired many studies. In this work, we analyze the number of persons moving from one city to another, regardless of the transportation or reason. Due to the large number of variables we can use to describe an urban region, models can be arbitrarily complex. The simplest approaches, the Gravity model[1] and the Radiation model[2], only use the distance and the population. However, more sophisticated deep learning approaches, such as the Deep Gravity model[4] can use 39 features.

The aim of this project is to use a symbolic regression method to recover closed-form mathematical models from the data, the Bayesian Machine Scientist(BMS)[**?**]. The only constraint we impose is that the variables that can appear in the model are the distance and the origin-destination populations. We use data from 6 states of the United States to train and test the models. Each state is trained with a sample of flows between a subset of cities and the same symbolic expression but with different parameters. To test the models we use the flows between a subset of cities different from the train set. To compare the performance of the BMS models, we train the mentioned reference models including a Random Forest with 39 features. To evaluate the predictions, we use the following metrics: Common Part of Commuter, Median Absolute Error, Median Relative Error, and Median Log-Ratio.

Our results show that the metrics Common Part of Commuters and Median Relative Error only capture the performance for high values of the flow. If we compare models, is difficult to distinguish a good performance from an overfitted model. We state relative metrics are more useful since we have data with different orders of magnitude. If we compare the results for the Median Log-Ratio, we see that BMS models with three variables perform similarly or better than complex models with up to 39 features.

[1] George Kingsley Zipf, *The P1 P2/D Hypothesis: On the Intercity Movement of Persons*, American Sociological Review **11** (1946).

[2] Simini, Filippo, González, Marta C., Maritan, Amos, Barabási, Albert-László, *A universal model for mobility and migration patterns*, Nature **484** (2012).

[3] Simini, Filippo, Barlacchi, Gianni, Luca, Massimilano, Pappalardo, Luca, *A Deep Gravity model for mobility flows generation*, Nature Communications **12** (2021).

[4] Roger Guimer, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel Miranda, Jordi Pallars, Marta Sales-Pardo , *A Bayesian machine scientist to aid in the solution of challenging scientific problems*, Science Advances **6** (2020).