nature communications

Article

https://doi.org/10.1038/s41467-025-56495-5

Human mobility is well described by closed-form gravity-like models learned automatically from data

Received: 14 December 2023

Accepted: 21 January 2025

Published online: 04 February 2025

Check for updates

Oriol Cabanas-Tirapu[1], Lluís Danús[1,2], Esteban Moro[3,4], Marta Sales-Pardo[1] & Roger Guimerà[1,5]

Modeling human mobility is critical to address questions in urban planning, sustainability, public health, and economic development. However, our understanding and ability to model flows between urban areas are still incomplete. At one end of the modeling spectrum we have gravity models, which are easy to interpret but provide modestly accurate predictions of flows. At the other end, we have machine learning models, with tens of features and thousands of parameters, which predict mobility more accurately than gravity models but do not provide clear insights on human behavior. Here, we show that simple machine-learned, closed-form models of mobility can predict mobility flows as accurately as complex machine learning models, and extrapolate better. Moreover, these models are simple and gravity-like, and can be interpreted similarly to standard gravity models. These models work for different datasets and at different scales, suggesting that they may capture the fundamental universal features of human mobility.

Accurate models of population mobility within and between municipalities are critical to address questions in urban planning and transportation engineering. Additionally, since municipalities are the main ground on which societies and cultures develop today, such mobility models are also instrumental in addressing global challenges in sustainability, public health, and economic development. Two main factors have driven recent interest in modeling human mobility patterns[1–5]. First, accurate models of human mobility could help identify transportation needs[6], allocate services and amenities (shopping, health, parks) more efficiently[7], or even understand and eventually alleviate problems like segregation[8], or epidemic spreading[9]. But, at the same time, models of human mobility can help identify the main behavioral components driving people to make large displacements to, for example, buy a new product, find a new house, or use physical activity spaces. Better behavioral models can help us implement more efficient policies to change people's behavior, rather than urban environments, in favor of more sustainable attitudes.

Despite these considerations, our understanding of the mobility flows within and between urban areas is still incomplete. One of the earliest and most fruitful attempts to model mobility flows between municipalities is the so-called gravity model[1]. This model assumes that mobility flows depend solely on the attractiveness or opportunities of the origin and destination municipalities (for which population is typically used as a proxy) and the geographical distance between them, in a fashion that is mathematically similar to Newton's law of gravitation. In its different incarnations and refined versions[4,10,11], the gravity model provides a simple phenomenological description of a very complex phenomenon. Because of this, while gravity models are not without their limitations, they are very often used in urban design, transportation, or even commercial applications. Recently, machine learning and deep learning algorithms have been proposed, extending the ideas underlying gravity models; they incorporate many other features besides the populations of the origin and destination municipalities and their distance[12–14]. Although those sophisticated machine learning tools are more accurate at predicting flows between urban

[1]Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain. [2]Annenberg School for Communication, University of Pennsylvania, Philadelphia, PA, USA. [3]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Network Science Institute, Northeastern University, Boston, MA, USA. [5]ICREA, Barcelona, Catalonia, Spain. e-mail: marta.sales@urv.cat; roger.guimera@urv.cat

areas, they lack interpretability and analytical tractability, and are hard to adapt to new contexts.

Given the reasonable success of simple gravity models at explaining human mobility flows, here we investigate the fundamental question of whether we really need models that are much more complex than the gravity law to delve deeper into the essence of urban mobility. Unlike other behavioral models, gravity mobility models are phenomenological. Because of their lack of precise theoretical underpinnings, their predictive ability depends on the exact functional specification of the dependency of the mobility flows on the model features; that is, the mathematical dependency on origin and destination populations and distance. Here, we leverage recent developments in Bayesian symbolic regression to obtain closed-form, interpretable models[15] of mobility from data in a principled and automatic fashion[16–18].

We systematically compare the performance at predicting mobility flows of simple gravity models, complex machine learning and deep-learning methods, and closed-form, interpretable models obtained through Bayesian symbolic regression (Fig. 1). We find that the Bayesian symbolic regression approach yields simple models that are as accurate as the best machine learning approaches, and extrapolate better to out-of-sample regions. Our approach is able to learn accurate models that, like gravity models, solely take into account the origin and destination populations and the geographical distance between them. Importantly, the learned models are gravity-like in their mathematical dependencies on populations and distance. We also show that closed-form gravity-like models with non-population features are similarly predictive. Furthermore, exploration of the relationship between the contribution of the populations of municipalities (or other non-population features) and their relative distance reveals common patterns in all the datasets, which suggests a close to universal relationship between mobility flows and these variables.

Results

A Bayesian machine scientist learns closed-form mathematical models from mobility data

We aim to determine whether it is possible to model mobility flows by means of closed-form mathematical models that are interpretable like gravity models, and as predictive as (non-interpretable) machine learning models such as the deep gravity model[12]. To automatically learn such closed-form models from data, we use the so-called Bayesian machine scientist (BMS)[16]. Given a dataset D, the BMS samples closed-form mathematical models from the posterior distribution $p(M|D)$, which gives the probability that a given model M is the true generating model given the data (Methods). The BMS is guaranteed to asymptotically identify the true generating model, if one exists. Additionally, when the data are truly generated from a closed-form model, the BMS has been shown to make quasi-optimal predictions for unobserved data[18].

We consider as our main dataset the set of flows T_{od} between origin o and destination d municipalities in six states in the USA (New York, Massachusetts, California, Florida, Washington, and Texas; see Data). The BMS is fed with D, a set of 1000 flows within each state, and samples closed-form models from $p(M|D)$ using Markov chain Monte Carlo[16] (MCMC; Methods and Fig. S8). Rather than sampling different models for each state, the same models are used for all states (see Methods and ref. 17), which amounts to assuming that mobility patterns arise from general mechanisms that are not state-dependent. In the spirit of gravity models, however, each state is allowed to have different parameter values. Model sampling yields an ensemble of hundreds of different state-independent, closed-form models for the flows T_{od} such as, for example,

$$\log T_{od} = A\left(1+\frac{B\left(\left(m_d+C\right)\left(m_o+D\right)\right)^{\beta}}{d_{od}}\right)^{\xi} \quad \text{or} \quad \log T_{od} = \log\left[A\left(\frac{B\left(m_d m_o + C m_d + D\right)}{d_{od}^{\alpha}}+1\right)^{\gamma}\right] \tag{1}$$

where $m_{o/d}$ is the population of the origin/destination municipalities, d_{od} is the distance between them, and A, B, C, D, β, and ξ are model parameters. These models are able to make predictions of test flows (not seen by the BMS during training, and with origins and destinations also not seen during training) that follow real values over several orders of magnitude (Figs. 1 and 2M–R). In what follows, we analyze in more depth this ensemble of models and its predictive abilities, vis-a-vis gravity models and machine learning models such as the deep gravity model. Later, we test the ability of the BMS approach to

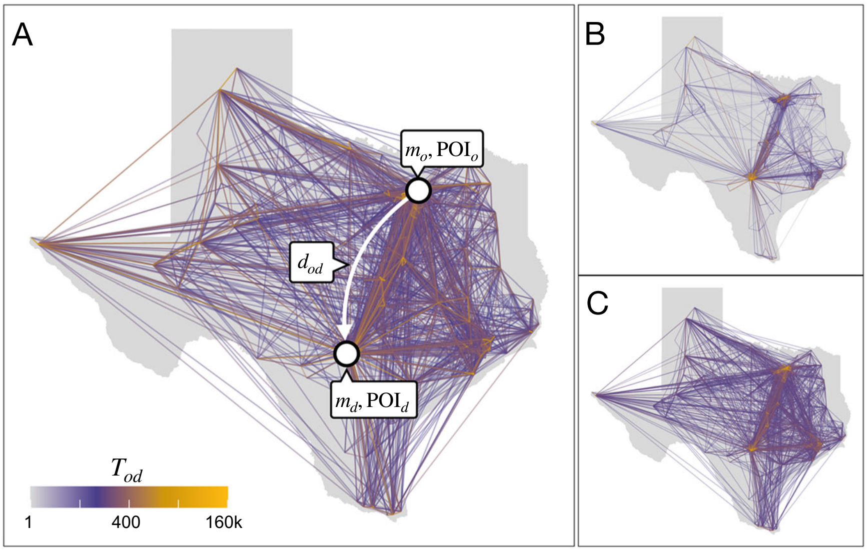


Fig. 1 | Modeling approaches for mobility flows between test municipalities in Texas, USA. A Real mobility flows between municipalities (T_{od}) in the test set in Texas, US (Methods). For each flow, we consider origin o and destination d features, such as population $m_{o/d}$, aggregate statistics about points of interest (POI), and the distance between them. **B** Flows predicted by the deep gravity model[12], which uses a total of 39 features from origin and destination. **C** Flows predicted by the closed-form, median predictive model (Fig. 6B) identified by the Bayesian machine scientist (BMS; see text). This model only uses the population of origin and destination, as well as the distance between them.

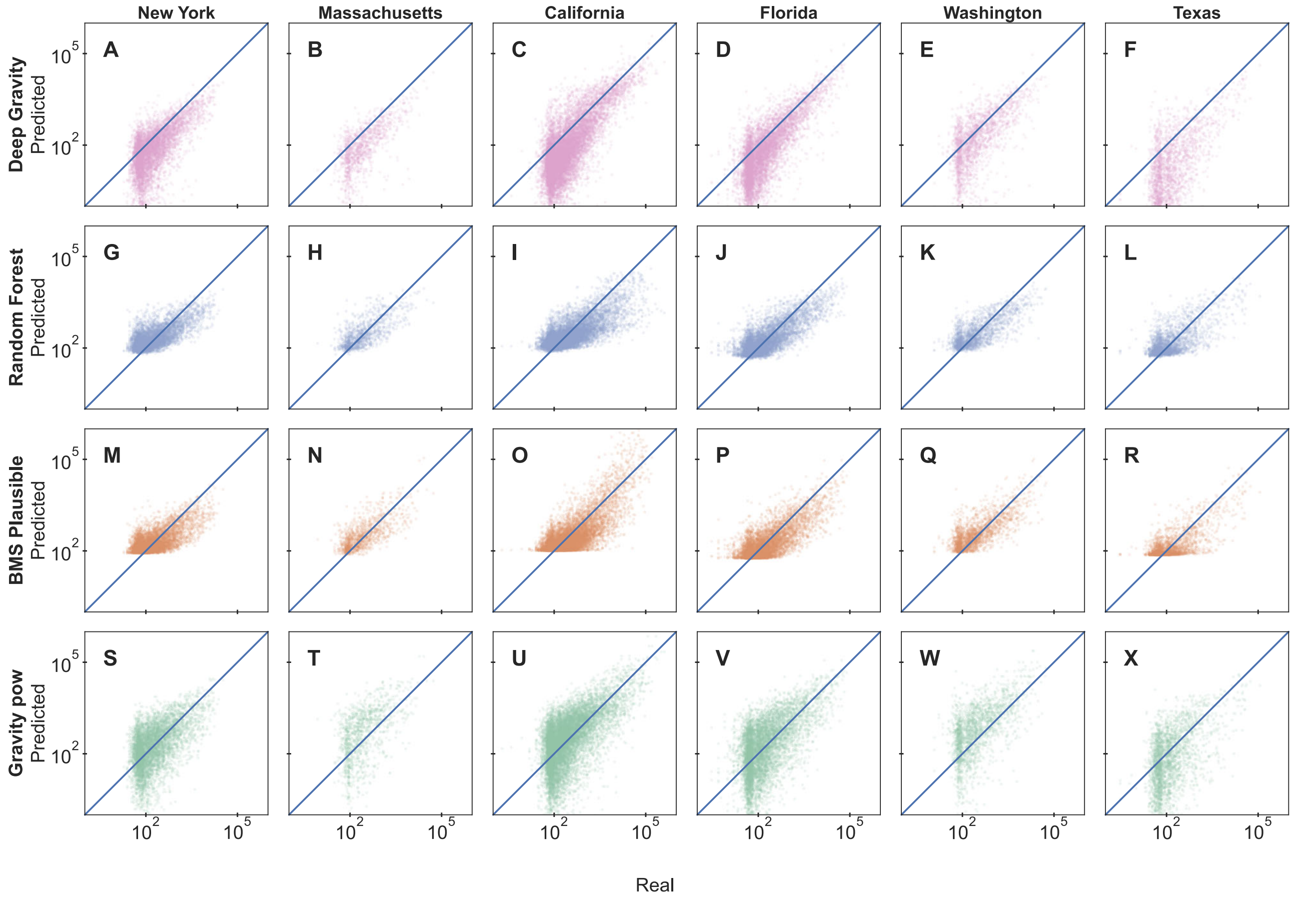


Fig. 2 | Model predictions of flows between municipalities. Each panel shows, in logarithmic scale, the scatter plot of predicted flows between municipalities versus the corresponding real flows, for different states in the US (columns). Plots show results for test data for different models (rows): **A**–**F** the Deep Gravity model, (**G**–**L**) a Random Forest regressor, (**M**–**R**) the most plausible model sampled by the Bayesian machine scientist, and (**S**–**X**) a gravity model in its power-law version. Supplementary Fig. S1 shows scatter plots for the full set of models we consider (see Methods for a complete description of the models and their parameters). Different models capture flows at different scales.

produce accurate models at different length scales (namely, to model flows between fixed-size tiles[12], as opposed to municipalities); and of the models identified for the original six states to make out-of-sample predictions on six different states (Georgia, Illinois, Michigan, North Carolina, Ohio, and Pennsylvania).

Different models capture flows at different scales

In order to compare the ability of modeling approaches to describe mobility data, one needs a model selection criterion. In probabilistic terms, selecting the best model amounts to selecting the most plausible model, that is, the model that has the highest probability $p(M|D)$ of being the true generating model given the observed data; or, equivalently, the model with the shortest description length (Eqs. (3)–(5)). However, this criterion is not always applicable in practice, because often it is not possible to compute the description length of a model, as it happens for deep learning and most other machine learning models.

Alternatively, one can measure performance at certain predictive tasks[19], which is the approach typically taken in mobility modeling studies and that we take here. Specifically, for each of the six training states (New York, Massachusetts, California, Florida, Washington, and Texas), we follow ref. 12 and split municipalities into two sets. Flows between municipalities in the first set comprise the training set, and flows between municipalities in the second set comprise the test set (see Tables 1 and 2). To make sure that all states carry a comparable weight in the training data D, we select the same number of flows from each state (1000 flows, as we are limited by the state with the fewest municipalities; Methods). By building the training and test sets in this way[12], all the information about the municipalities in the test set, their characteristics, and the distances between them is completely new to the trained algorithm. Additionally, since the geographic location of municipalities is not available to any of the algorithms, potential similarities between close locations in the train and test datasets are not patterns that can be learned by any of the models.

We compare the closed-form mobility models identified by the BMS to two alternative approaches. First, we consider gravity models, in which mobility flows are directly proportional to the product of masses (that is, populations) at the origin and destination, and inversely proportional to the distance between them. These approaches include traditional gravity models[1], as well as the closely related radiation model[4]. Second, we consider machine learning approaches that, besides considering population and distance between municipalities, also consider additional characteristics of municipalities, such as the density of shops, entertainment venues, or educational facilities (Methods). Specifically, we consider a random forest regression model[20] and the deep gravity model[12].

From the ensemble of closed-form models sampled by the BMS, we analyze (Methods): (i) the most plausible (minimum description length) model found by the BMS; (ii) the median of the ensemble of models sampled by the BMS (which is the optimal predictor); and (iii) the single model in the ensemble of sampled models whose predictions are closest to the ensemble median, which we call the median predictive model.

Table 1 | Train dataset

State	Entries	Municipalities	Flow		Distance (Km)		Population	
			Min	Max	Min	Max	Min	Max
New York	1000	217	26	122,991	1.83	489.80	185	8.80×10^6
Massachusetts	1000	75	19	6,6946	2.34	205.51	1029	6.76×10^5
California	1000	260	14	91,267	2.12	1078.21	237	1.01×10^6
Florida	1000	223	7	78,316	2.54	607.50	251	9.50×10^5
Washington	1000	107	20	71,796	2.17	467.16	20	2.29×10^5
Texas	1000	177	21	192,660	1.84	975.16	173	2.30×10^6

A number of points and municipalities in the train set for each state was obtained from a random sample of 1000 points of the original train fold. We also detail the lowest and largest flow, distance, and municipality population.

Table 2 | Test dataset

State	Entries	Municipalities	Flow		Distance (Km)		Population	
			Min	Max	Min	Max	Min	Max
New York	5952	249	20	35,063	0.77	531.39	361	2.12×10^5
Massachusetts	1180	75	9	55,499	2.16	267.24	1,517	2.07×10^5
California	11,727	319	2	41,6696	1.09	1,084.34	129	3.9×10^6
Florida	7092	245	3	79,624	1.14	875.82	78	4.42×10^5
Washington	2083	109	12	54,245	1.48	431.53	487	7.37×10^5
Texas	2782	192	3	112,865	2.46	1,196.33	106	1.43×10^6

Number of points and municipalities in the test set for each state. We also detail the lowest and largest flow, distance, and municipality population.

In Fig. 2, we show the predicted flows versus the real flows in the test set (see Supplementary Fig. S1 for results for additional models). Whereas all models are predictive, we find that different models are differently capable of describing mobility flows of different orders of magnitude. For instance, gravity-like models (Fig. 2S–X and Supplementary Figs. S1 and S2) are typically good at capturing the behavior of large flows, but not of small flows. Indeed, for small flows (less than around 100 commuters) these approaches tend to underestimate flows, in some cases by several orders of magnitude, and even predict flows smaller than 1 person (Supplementary Fig. S2). This is also the case for the deep gravity model, which again under-predicts small flows (Figs. 1 and 2A–F). By contrast, neither the random forest nor the BMS suffers from this caveat, and both capture the whole range of flows more consistently and without large systematic deviations (Figs. 1 and 2G–R).

Simple closed-form models are as accurate as the best machine-learning models on training states

Next, we quantify the performance of the models at the task of predicting unobserved flows. To that end, and considering the qualitative results in the previous section, we compute several complementary performance metrics. First, we consider the common part of commuters (CPC; see Methods), which is a usual choice in the mobility literature[12]. The CPC measures the overlap between predicted and observed flows, and can take values from 0 to 1; the larger the CPC, the better the predictions. Despite its popularity, this metric favors models that predict the larger flows well, but overlooks errors in small flows (Fig. S3A–F). Since mobility flows typically span several orders of magnitude (Fig. 2), models with the larger CPC are not necessarily the best models for the whole range of flows.

To have metrics of performance that cover the whole range of flows, we consider, in addition, and complementary to CPC, the absolute error, the absolute relative error, and the absolute log ratio (Methods, Fig. 3, Supplementary Fig. S3). For each of these metrics, and to avoid the disproportionate influence of singular large errors (especially for non-relative quantities such as the absolute error), we always show the whole distribution of error values (as a boxplot), and use the median value to compare models (Fig. S3, Supplementary Table S6); the lower the median, the better the performance of the model. Note that these metrics highlight different aspects of the prediction. Absolute errors are correlated with the magnitude of the flow we are trying to predict so that errors are typically larger for larger flows. Because of this, and similar to the CPC, average absolute errors are very sensitive to the errors in predicting large flows but not to errors in small flows. For the same reason, median values of the absolute error typically reflect errors in performance for typical flow values and do not reflect the ability of a model to predict values in the whole range of flows.

The absolute relative error and the absolute log-ratio do take into account the effect of the magnitude of the flow, and therefore are more informative of the global behavior of a model when the range of flows spans several orders of magnitude (Fig. S3M–X). An issue with the relative error is that while it penalizes over-prediction, it does not penalize under-prediction; in the extreme case in which the predicted flow equals zero and the real flow is larger than zero, the relative error is equal to one. As a result, distributions for relative errors in gravity-like models and the deep gravity model, in which small flows are under-predicted, are centered around 1 (Supplementary Figs. S2, S3M–R). By contrast, the absolute log-ratio has the property that over- and under-prediction are equally penalized (in a logarithmic scale), that is, predicting the real flow multiplied or divided by the same factor results in same absolute log-ratio. This metric, therefore, captures the ability of a model to predict flows in any range of values (Supplementary Fig. S3S–X).

Using these metrics, we compare the different modeling approaches (Fig. 3 and Supplementary Fig. S1). The first conclusion from this comparison is that gravity models, including the radiation model, are never the best performers; for all states and metrics considered, there is always at least one other model that performs better. This is not surprising, since these models are simple and highly

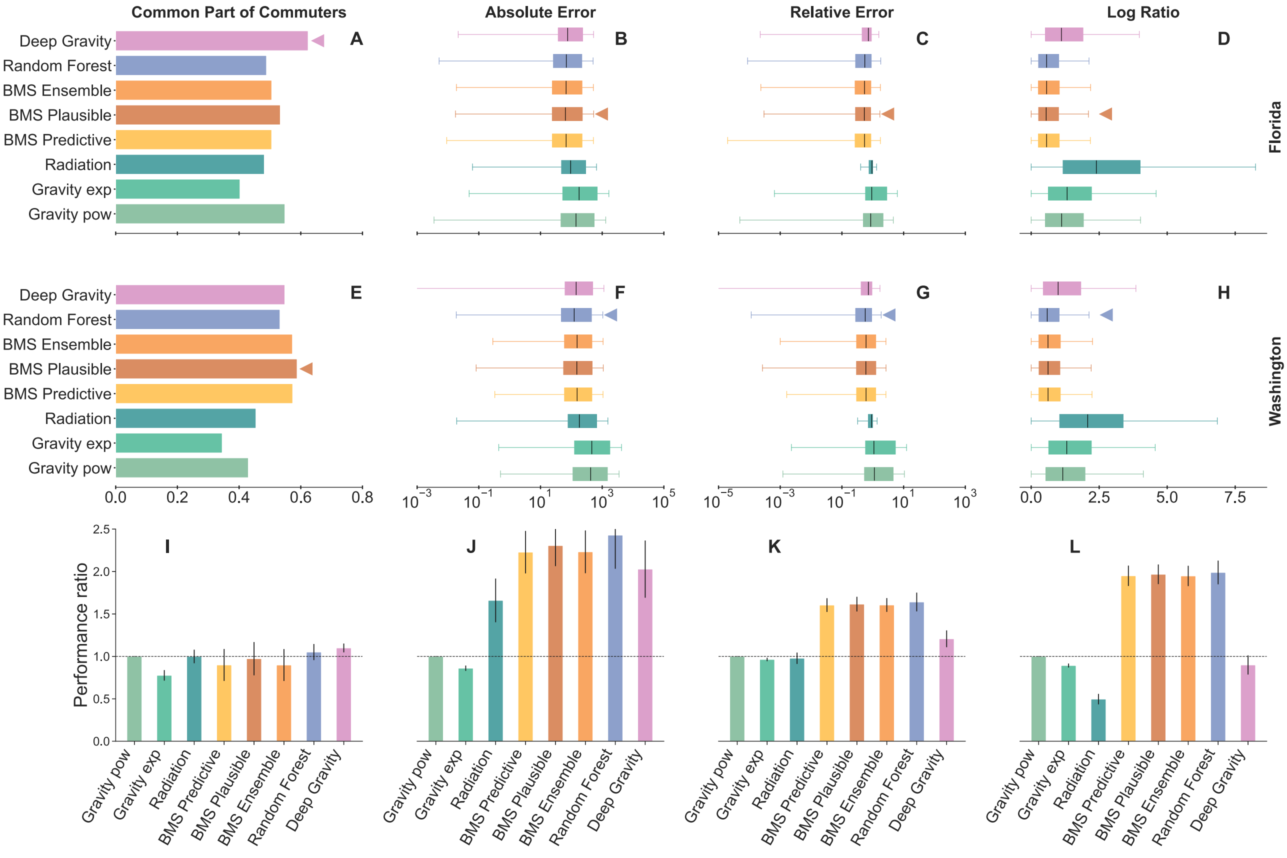


Fig. 3 | Model performance at predicting test flows between municipalities in training states. A–H For each model prediction for two representative states (Florida and Washington; see Fig. S3 for the remaining states), we assess model performance using four different metrics: **A**, **E** Common part of commuters; **B**, **F** Absolute error; **C**, **G** Absolute relative error; **D**, **H** Absolute log-ratio. The common part of commuters (CPC) is a global metric. Thus, we have a single value for each metric. For the other three metrics, we show the median, 50% confidence interval (box), and 95% confidence interval (whiskers). Triangles (◀) indicate the best-performing model for each metric (largest CPC or lowest median). See Methods and Text for the definition and discussion of the different metrics. **I–L** Summary of performance over test flows in the six training states. The performance ratio is defined with respect to the performance of the gravity model with power law decaying dependence on the distance (Gravity pow); values larger than 1 correspond, for all metrics (including CPC), to performance above the Gravity pow model, whereas values smaller than 1 indicate worse performance. Error bars indicate 95% confidence intervals for the means over states. (See Supplementary Table S6 for numerical values for all individual states, as well as summary statistics.) Overall, all models perform similarly in terms of CPC, whereas BMS models and random forests (and deep gravity for absolute error) perform significantly better in the other statistics.

stylized, and they have already been shown to make less accurate predictions than deep gravity models[12].

Perhaps more surprisingly, we find that closed-form mathematical models obtained by the BMS perform, overall, comparably to random forest models and better than deep gravity models. Indeed, the most plausible model (BMS Plausible) performs better than deep gravity for 21 of the 24 comparisons we can make (four metrics on each of the six states), whereas it performs better than random forests in 13/24 comparisons and worse in the remaining 11/24 (Fig. S3). On average, over the six states, BMS models perform indistinguishably from random forests on all quality metrics, and better than deep gravity models in terms of relative error and log ratio (Fig. 3I–L). Remarkably, deep gravity models are only best performers in terms of CPC, and even they are not significantly better than BMS models or random forests, a result that is consistent with previous findings[13].

When considering how each model performs for flows in specific ranges (Fig. 4 and Supplementary Fig. S4), we find that BMS models, similar to random forest models, are particularly good at modeling flows in the range that is the most common in the data.

Taking into account that both random forest and deep gravity models use many more features for their predictions (39 features in total, in contrast to the three features used by gravity models and the closed-form models identified by the BMS, namely, origin and destination population, and origin-destination distance), we conclude that the symbolic regression approach using the BMS yields parsimonious models of human mobility flows between municipalities. Closed-form models obtained by the BMS also compare well to the alternatives in terms of the fairness of their predictions[21] (Supplementary Fig. S5). In particular, we find that the random forest and the models identified by the BMS are, overall, the most consistent models across states in terms of the fairness of their predictions.

Closed-form models also describe flows at shorter scales

So far, we have analyzed within-state flows between municipalities, of any size and at any range of distances. However, it may be that the geographical, economic, and demographic characteristics of smaller areas become relevant when modeling flows at shorter distances (for example, within neighborhoods or adjacent towns in large metropolitan areas). To elucidate to what extent different modeling approaches can accommodate into such short-distance flows, we adopt the framework used in ref. 12—we divide the state of New York in small tiles of 25 × 25 km^2, and consider the flows between census tracts within each tile[12]. We use 50% of the tiles for training the different models, and then test the flows between the remaining 50% of the tiles. For this experiment, we find that regardless of the metric used, closed-form models identified by the BMS are always more accurate than machine

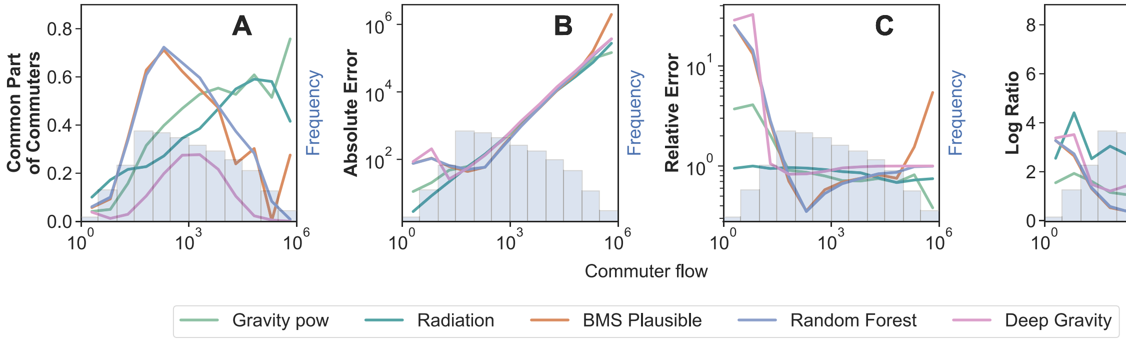


Fig. 4 | Performance for different flow ranges. The observed flows between municipalities span six orders of magnitude and are distributed, over the six states in our main dataset, as shown by the histogram. We measure the performance of the different models at predicting flows within each bin of flows: **A**, CPC; **B**, absolute error; **C**, absolute relative error; **D** absolute log ratio. The best-performing models (random forests and those identified by the BMS) tend to fit the data particularly well (higher CPC and lower error) when flows have the most common magnitudes.

learning and gravity models (Fig. 5). Our results thus indicate that simple closed-form models that just consider populations of municipalities/census tracts and the distances between them provide better descriptions of mobility flows than complex models that take many more features into account, and have many more parameters, also for flows at short distances. This result is consistent with previous work suggesting than, also at these scales, adding many features to mobility models barely improves prediction[14].

The identified closed-form models are gravity-like and generalize well

Our analysis indicates that the BMS is able to find closed-form mathematical models that solely consider the populations of the origin and destination and the geographic distance between them; and that these models provide predictions of mobility flows that are as accurate as the most accurate machine learning models. Here, we investigate whether, besides being predictive, these closed-form models are also interpretable and insightful, and if they generalize well to other contexts.

We start by noting, once more, that the BMS samples hundreds of models and that, overall, they all perform well, which shows that there are many different models that can describe the data. Such a set of models is sometimes called a Rashomon set[15]. Then, the relevant question is whether these models share any common defining properties that could explain why they describe mobility flows accurately. To elucidate this question, we consider a collection of 105 models sampled by the BMS (see Supplementary Information). We notice that many of these models contain gravity-like terms, that is, terms that depend on a product of the origin and destination populations (perhaps shifted by a certain amount), and inversely on a growing function of the geographic distance between origin and destination. To quantify this observation, and based on this definition of gravity-like terms, we manually classified the 105 models into three groups: those having only gravity-like terms, those having gravity-like terms and some other (multiplying or additive) terms, and those not having any gravity-like terms. We find that 17% of the sampled models are purely gravity-like and 70% contain gravity-like terms as well as other terms; only 13% do not contain any gravity-like terms at all. This is remarkable because the BMS has not received any input about the particular shape that models should take, which suggests that the regularities in the data are well-described by this general class of models and justifies the historical use of gravity models.

Next, we analyze in more detail two particularly relevant closed-form models identified by the BMS (Fig. 6): (i) the most plausible model, that is, the model that has the highest probability $p(M|D)$ given the data (or, equivalently, the shortest description length $\mathcal{L}(M,D)$) among all those sampled by the BMS (Methods); (ii) the median predictive model, that is, the closed-form model whose predictions for unobserved data are closest to the median prediction of the whole ensemble of sampled models (Methods). Formally, there are marked mathematical differences between both models. In particular, the most plausible model is an exponential model for the flows, while the median predictive model is a power-law model for the flows. However, the two models have relevant properties in common. First, both models are also gravity-like models, like the majority of models sampled by the BMS. Second, origin and destination do not necessarily play a symmetric role, which allows the model to accommodate non-symmetric flows in contrast to typical gravity models which do not allow for this possibility. Indeed, an inspection of the parameters shows that in some states, such as New York or Texas, flows are much more symmetric than in others, such as Florida and Massachusetts. Third, the relative contribution of the mass product with respect to the geographical distance is quite consistent—in both models, we find a mathematically equivalent dependency on the ratio $(m_d m_o)/d^e$, where $e = 1/\beta$ in the most plausible model (Fig. 6A), and $e = \alpha$ in the median predictive model (Fig. 6B). We find that, for a given state, α close to $1/\beta$ suggesting that this relationship is to a large extent model-independent (Fig. 6C). Furthermore, we find that the state-to-state variability is relatively small since all exponents fall within a close range, which suggests that reasonable models for mobility flows are gravity-like models with specific constraints in the relationship between the contributions of the mass product and the geographical distance.

To conclude our analysis of the most plausible and median predictive models, we study to what extent they generalize to out-of-sample scenarios (Fig. 6D, E). To that end, we use mobility flows between municipalities in six new states: Georgia, Illinois, Michigan, North Carolina, Ohio, and Pennsylvania. For each new state, we again split municipalities in train and test set, and use the train set to fit model parameters and the test set to measure the models' generalization ability. As we show in Fig. 6E, the original models are as accurate in the new states as in the original states, confirming that both models generalize well (see Supplementary Fig. S6 for model parameters in the new states). We also compare the generalization performance of the BMS models to that of the gravity and random forest models (Fig. 6D). We find that, at this generalization task, BMS models perform significantly better than random forests for three out of four quality metrics, and indistinguishably in the fourth. It is worth noting that because random forests are non-parametric, in this case, we cannot refit the parameters of the models to the new states; rather, we predict each flow in the new states by averaging over the predictions of the six original random forest models. Doing the same with BMS models instead of refitting model parameters for the new states, still leads to flow predictions in the new states that are

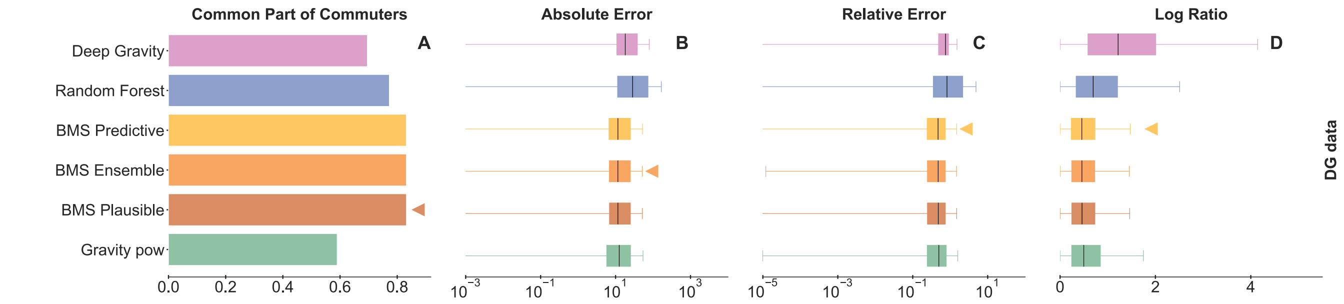


Fig. 5 | Model performance at predicting flows at small distances. We consider the data used by Simini et al.[12] about mobility flows between census tracts within small geographical (25 × 25 km^2) regions in NY state. We evaluate predictions over test data using the same metrics as in Fig. 3: **A** Common part of commuters; **B** Absolute error; **C** Absolute relative error; **D** Absolute log-ratio. In each panel, triangles (◂) indicate the best-performing model according to the corresponding metric; BMS models are the best-performing in all metrics. See Methods and text, for a complete description of the metrics, the models, and the data.

significantly more accurate than those of random forest models (supplementary Fig. S7).

Gravity-like models with non-population features are also highly predictive

In the preceding sections, we have shown that gravity-like models based only on population and distance are as predictive as complex machine learning models with population and non-population features, and extrapolate better. However, previous research suggests that, in general, non-population features add predictive power to models[12], which is consistent with our current understanding of the diverse factors that drive human mobility.

To clarify this apparent paradox, we end by investigating closed-form models with population and non-population features (Fig. 7). We start by noting that, at the level of municipalities and in contrast to smaller scales, all features are highly correlated (Fig. 7A–C), which explains why population and distance, alone, already yield highly predictive models. In any case, we feed the BMS with the 39 features used by deep gravity and random forest models, to obtain new closed-form models with those features. We find that most of the features are rarely used in the sampled expressions, but a few are (in particular, the number of food points at the destination, and the main road lines at origin and destination).

The most plausible model identified by the BMS is, in this case,

$$\log T_{od} = \left[c_1 + \frac{c_2\left(c_3 + \left(c_2 F_d + M_d\right)\left(c_4 + M_o + R_o\right)\right) \exp\left(c_5^{R_o}\right)}{d^{\alpha}} \right]^{\gamma} \tag{2}$$

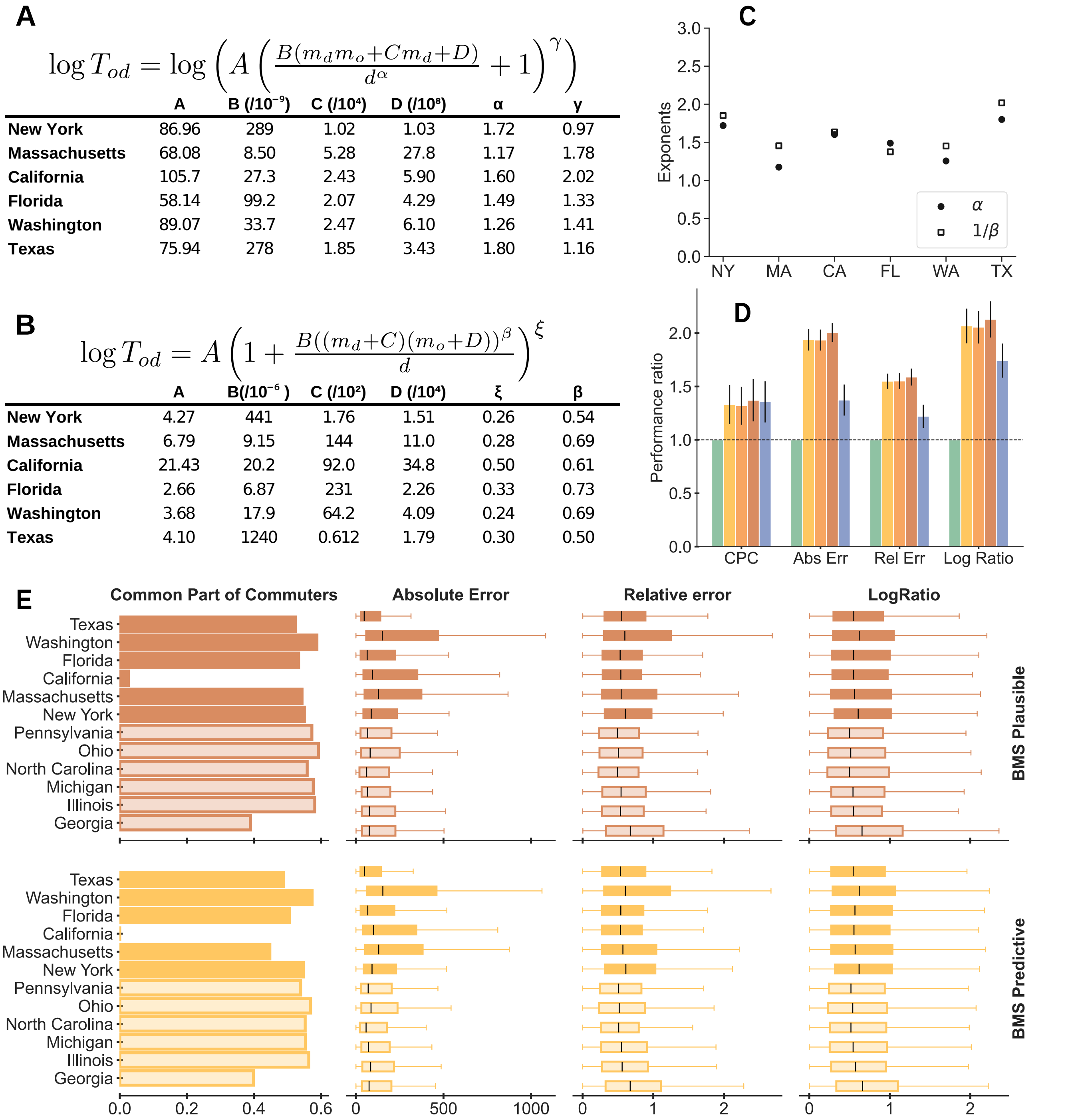
where F_d is the number of food points at the destination, $M_{o/d}$ is the length of main road lines at the origin and destination, respectively, and R_o is the number of retail points at the origin. We find that this model is often slightly more predictive than the population-only gravity-like models discussed in the previous sections, although the differences are not statistically significant. This is true for the six original states (Fig. 7C–F) as well as for the six new states not seen by the BMS (Fig. 7G–J). Last but not least, we note that this model is also gravity-like in that it contains a term with the product of an origin-only factor and a destination-only factor, divided by an increasing function of the distance. Thus, formally, this model is similar to population-only models in previous sections.

Discussion

Understanding human mobility is critical to address questions in urban planning and transportation, as well as global challenges in sustainability, public health, and economic development. Traditionally, mobility flows have been modeled using simple gravity models, which are conceptually simple and easy to interpret, but have limited predictive power. Recently, deep learning models have been proposed as an alternative; whereas these models are consistently and significantly more predictive than gravity models, they are not interpretable and provide little insight into human behavior. Here, we have shown that automated equation discovery approaches lead to parsimonious closed-form models that combine the most desirable aspects of both approaches—the simplicity of gravity models and predictive power as high as the most accurate machine learning models and even higher when it comes to generalizing.

We argue that two factors contribute to the success of the BMS at identifying parsimonious models. First, the probabilistic approach underlying the BMS deals quasi-optimally with sparse and noisy data[18]. Mobility datasets are available nowadays, and more will be available in the future[22], but the number of municipalities in a state or even a country is limited, so the resulting training sets are constrained by definition. This is in contrast with LLMs or other areas where deep learning excels, and where the size of training sets can grow virtually without limits. Second, automated equation discovery works best when the data can be described by relatively simple models. This seems to be the case in the context of mobility. Indeed, the models we identified are gravity-like in that they are increasing functions of a certain product of populations of origin and destination, and decreasing functions of the distance between them. While the ratio between the population and the distance terms is, in principle, model and data-dependent, in the datasets we explore the ratio is roughly constant and dataset-independent.

Individual mobility depends critically on the urban environment, personal preferences, commuting patterns, and accessibility to transportation and amenities. Thus, modeling mobility at an individual or small spatial scale might require more complicated models that account for routes, the purpose of the trip, points of interest, or even the demographic traits of individuals[12,21,23]. However, our results show that by aggregating mobility at a larger spatial scale, the movements of millions of people can be described parsimoniously by simple fully explainable models that do not depend on the microscopic characteristics of the origin, destination, or route taken. This is because the randomness and variability inherent in individual behaviors tend to cancel out when looked at collectively, revealing underlying trends and movements that are driven by the shared needs of large populations, and the structure of the built environment. Our results show, therefore, that the aggregated flows in human mobility can be seen as an emergent and universal property of the complex system of individual movements. More broadly, our results showcase the potential of using machine scientists[16,24,25] to automate the process of finding similar phenomenological closed-form models from data; and to use these models to gain insight into the relevant variables and mechanisms to describe other complex phenomena.



A

$$\log T_{od} = \log\left(A\left(\frac{B(m_d m_o + C m_d + D)}{d^{\alpha}} + 1\right)^{\gamma}\right)$$

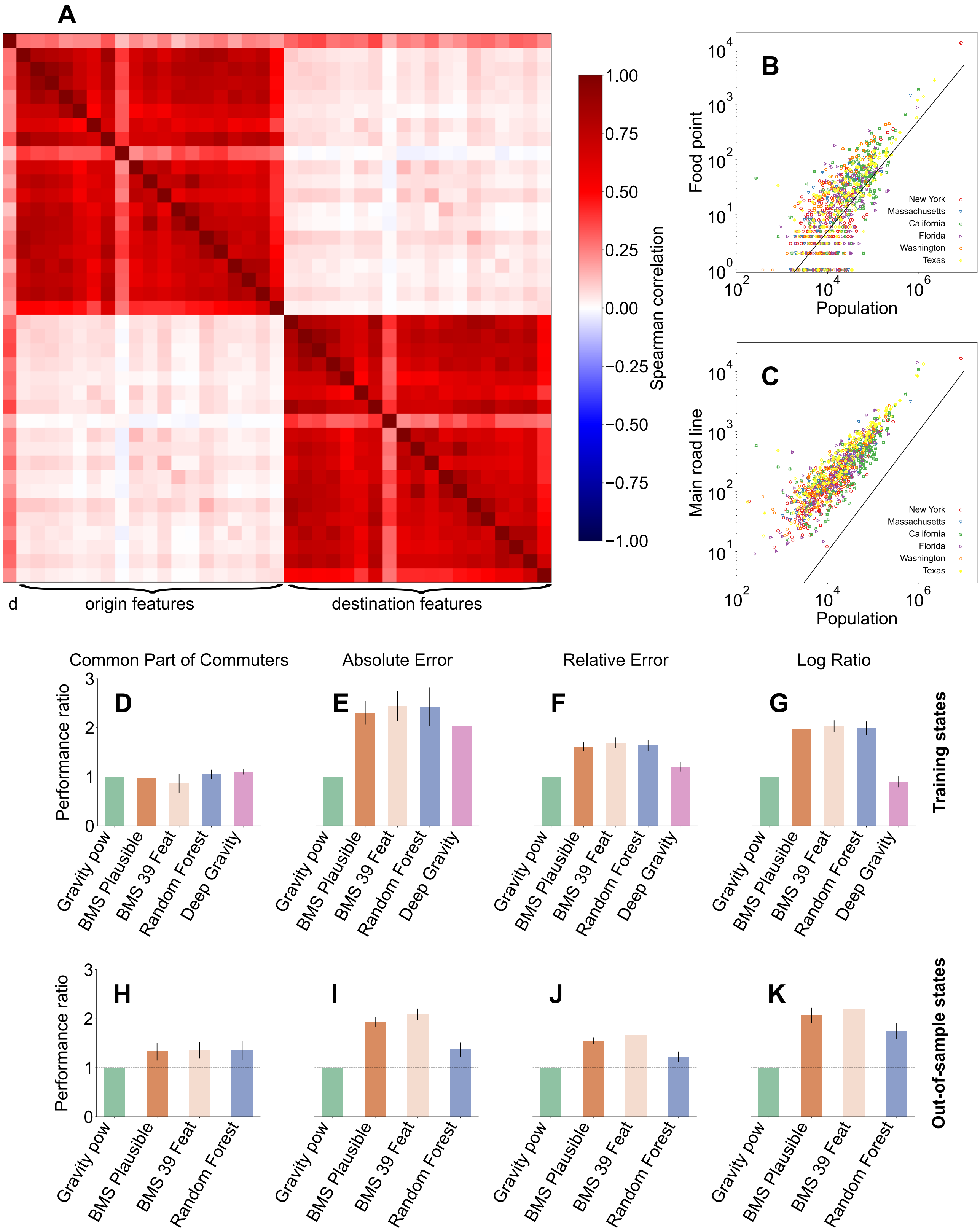
	A	B (/10^{-9})	C (/10^{4})	D (/10^{8})	α	γ
New York	86.96	289	1.02	1.03	1.72	0.97
Massachusetts	68.08	8.50	5.28	27.8	1.17	1.78
California	105.7	27.3	2.43	5.90	1.60	2.02
Florida	58.14	99.2	2.07	4.29	1.49	1.33
Washington	89.07	33.7	2.47	6.10	1.26	1.41
Texas	75.94	278	1.85	3.43	1.80	1.16

B

$$\log T_{od} = A\left(1 + \frac{B((m_d + C)(m_o + D))^{\beta}}{d}\right)^{\xi}$$

	A	B(/10^{-6})	C (/10^{2})	D (/10^{4})	ξ	β
New York	4.27	441	1.76	1.51	0.26	0.54
Massachusetts	6.79	9.15	144	11.0	0.28	0.69
California	21.43	20.2	92.0	34.8	0.50	0.61
Florida	2.66	6.87	231	2.26	0.33	0.73
Washington	3.68	17.9	64.2	4.09	0.24	0.69
Texas	4.10	1240	0.612	1.79	0.30	0.50

Fig. 6 | Closed-form models for mobility flows. We ran the Bayesian machine scientist (BMS) with a training set of 1000 points and three features: origin and destination populations and the distance between them. We used 5 independent Markov chains of 12,000 Monte Carlo steps each. **A** Minimum description length model (Methods) for the logarithm of the data, where d is the inter-municipality distance, m_o is the origin population, and m_d is the destination population. In the table, we show the fitting parameters for each state of the training data. **B** Median predictive model (Methods) for the logarithm of the data. As before, d is the inter-municipality distance, m_o is the origin population, and m_d is the destination population. In the table, we show the fitting parameters for each state of the training data. **C** Ratio between the distance exponent and the population exponent. Round-filled points are obtained from the most plausible model, and empty square points are obtained from the median predictive model. **D** Relative improvement of each model and metric for the out-of-sample states. We average the relative improvement across the out-of-sample-states. As in Fig. 3I–L, higher values of the performance ratio are always better (also for CPC). BMS models perform significantly better than all other algorithms in three out of four metrics. **E** We fit the BMS Plausible and BMS Predictive model parameters for the out-of-sample states. We compute the metric using the observed values from the test set and the predicted values of the model using the corresponding set of fitted parameters for each state. Filled bars show the training states and empty bars show the out-of-sample states. Performance in out-of-sample states is similar to training states, confirming that BMS models generalize well.



Methods

Mobility data between municipalities in the US

We collected weekly flows between United States census tracts for a week in January 2019 (2019/01/07 to 2019/01/13) and a week in March 2019 (2019/03/04 to 2019/03/9)[26]. The data consist of anonymous mobile data trajectories extracted from location-based service apps, regardless of the transportation method, and corresponding to the daily movement of millions of mobile phone devices in the US (see ref. 26 for a full description of the data). The dataset contains the geographical identifier (GEO ID) for both origin and destination census tracts as well as their corresponding geographical coordinates, the estimated number of visitors detected by SafeGraph, and the estimated population flows inferred from the number of visitors. The datasets are available at GeoDS.

Fig. 7 | Closed-form models with non-population features. A Spearman's rank correlation among all pairs of features (distance, 19 origin features, and 19 destination features), for all pairs of municipalities in the training states: New York, Massachusetts, California, Florida, Washington, and Texas. **B**, **C** Population of the destination versus: **B** number of food points at the destination; **C** main road lines at the destination. In both cases, the straight line represents a linear relationship; this is not a fit, and it is provided as a guide to the eye. **D**–**G** Performance ratio of models with respect to the gravity power-law model for the training states. As in Fig. 3, the performance ratio is defined so that values larger than 1 correspond, for all metrics (including CPC), to performance above the gravity power-law model, whereas values smaller than 1 indicate worse performance. Error bars indicate 95% confidence intervals for the means over states. The model labeled *BMS 39 Feat* corresponds to the most plausible model obtained by the BMS when trained with non-population features (Eq. (2)). **H**–**K** Same as **D**–**G**, but for out-of-sample states. The testing procedure is as in Fig. 6. In general, the *BMS 39 Feat* model performs slightly better than the most plausible population-only model, but the differences are not statistically significant.

The dataset was validated against ACS commuting flows and other datasets in the original publication by Kang et al., where it was found that there is a remarkable agreement (correlation above 90%) between the dataset and other administrative and commercial datasets[26]. SafeGraph data has also been validated and found to describe quite accurately flows and visits in the USA[27]. All of these validations are necessary because of the importance and challenges of getting reliable mobility data[22,28,29].

Data processing. In order to obtain the flows between municipalities from the data using census tract data, we first match municipalities (cities, towns and villages) with their corresponding census tracts. Then, the total flow between two municipalities A and B is calculated as the aggregate of flows between the set of census tracts municipality A is comprised of and the set municipality B is comprised of.

Specifically, we consider mobility datasets within six states in the US: New York, Massachusetts, California, Florida, Washington, and Texas for the training process. For validation, we also consider the following out-of-sample states: Georgia, Illinois, Michigan, North Carolina, Ohio, and Pennsylvania (See Supplementary Tables S1–S3). For each state, our data consist of the origin-destination municipality names, the flow, the distance, and the origin-destination populations and POI categories. We only consider municipalities with a non-zero population and pairs of municipalities with non-zero flow; we do not consider flows within the same municipality (see Supplementary Table S4 for details).

Information about municipalities and census tracts. We obtained shapefiles containing the geographical coordinates of the polygons delimiting census tracts and municipalities from United States Census Bureau. We also retrieved population data of each municipality using the GEO ID Data Commons.

Finally, using a local copy of Open Street Map (OSM) with Overpass API, we retrieved information about Points of Interest (POI) in each census tract (see Supplemantary Information and supplementary Table S5). We selected 18 categories of OSM elements that represent geographical, demographic and socio-economic features of the different municipalities.

Construction of train and test datasets. To speed up the training process, we train the models with the same random sample of 1000 points of the training fold except for the Deep Gravity model for which we have to use a lager number of data points for training.

Economic classification of municipalities. We retrieve the median income per capita of each municipality and state as of 2020 from Data Commons. Then, we classify each municipality with the label rich if the median income of the municipality is above the median income of the state, or poor if the median income per capita of the municipality is below the median income of the state.

Mobility data at small scales from Simini et al.[12]

The code available for the Deep Gravity model provides data at the level of census tracts for the New York State area. Mobility flows are obtained from the same dataset. In order to obtain the predictions of the Deep Gravity model for each individual trip and also the same dataset, we run the program and save, for each trip, the corresponding tile, the real value, the prediction and the variables. We store the train and test sets for the comparison with other methods.

Bayesian machine scientist

The BMS is a Bayesian approach to symbolic regression that estimates the plausibility of a closed-form mathematical model M given the observed data D as the posterior probability $p(M|D)$. Without loss of generality, this posterior can be written[16]

$$p(M|D) = \frac{\exp[-\mathcal{L}(M, D)]}{Z}\,, \tag{3}$$

where $\mathcal{L}(M, D)$ is the description length[30] of the model (and the data), and $Z = \sum_{M'} \exp\left[-\mathcal{L}(M', D)\right] = p(D)$ is the evidence. Here, we assume that the mobility flows are generated by the models M as

$$\log T_{od} = M(m_o, m_d, d_{od}; \theta) + \epsilon\,,$$

where θ are model parameters, and ϵ is an unbiased, Gaussian observational noise. Note that we model the logarithm of the flows rather than the flows themselves because, with flows spanning five orders of magnitude, the assumption of additive noise only makes sense for the logarithm. In this case, and within standard approximations[16,31], the description length can be approximated as

$$\mathcal{L}(M, D) = \frac{B(M, D)}{2} - \log p(M)\,, \tag{4}$$

where $B(M, D)$ is the Bayesian information criterion[31] (which is straightforward to calculate from the data), and $p(M)$ is a suitable prior distribution. Here, exactly as in ref. 16, we set $p(m)$ to be the maximum entropy distribution over models compatible with the frequencies of each elementary function observed in an empirical corpus of mathematical expressions. Note, in particular, that this prior does not give any preference to gravity-like models.

In this work, our goal is to simultaneously model flows from six different states in the US (that is, six different datasets). To this end, we use a multi-dataset approach[17] which consists in finding a unique closed-form model for multiple datasets. For each dataset, we allow model parameters to take different values[17]. For a single dataset $D = \{(y_i, \mathbf{x}_i)\}$, where $\{y_i\}$ is the set of observations and $\{\mathbf{x}_i\}$ is the set of feature values associated to each observation, the description length of a closed-form model M and the data is given by Eq. (4). In the case in which our data comprises K independent datasets $D = \{D_k,\ k = 1\ldots, K\} = \left\{\{(y_i^1, \mathbf{x}_i^1)\}, \ldots, \{(y_i^K, \mathbf{x}_i^K)\}\right\}$ that we want to model using a single model M, the description length is[17]

$$\mathcal{L}(M, D) = \frac{1}{2}\sum_k B(M, D_k) - \log p(M)\,. \tag{5}$$

The BMS represents closed-form models as labeled trees and uses Markov chain Monte Carlo (MCMC) to explore the space of closed-form mathematical models by sampling from the posterior distribution $p(M|D) \propto \exp(-\mathcal{L})$. In this work, we consider models for flows with: i) three independent variables (populations at origin and destination and geographical distance) and up to six parameters and II) 39 independent variables and up to 39 parameters.

The ensemble of sampled models allows to make predictions on the test set using three different approaches:

1. The most plausible model. This is the model with the shortest description length that the BMS is able to find.
2. The ensemble of models. Ensemble predictions are the optimal predictions since they correspond to fully integrating over model space. We estimate this integral by averaging over the predictions made by each one for the models we sample. Specifically, we perform five independent realizations of 12,000 MCMC steps (see supplementary Fig. S8 for typical traces of the MCMC sampling). We then collect a model every 100 steps within the last 2000 steps of the Markov chain to obtain an ensemble of 100 models.
3. The median predictive model. This is the model within the ensemble whose predictions are closest to the predictions of the ensemble as a whole.

Benchmark models

Gravity model. We consider the gravity model in its simplest form[1] in which the observed flow T_{ij} between municipalities (i,j) is a function of the populations m_i and m_j of the municipalities and the distance d_{ij} between them

$$T_{ij} = C\frac{m_i\, m_j}{f(d_{ij})}\,. \tag{6}$$

Here C is a scaling parameter and $f(d)$ is a function of the distance. Specifically, we consider two possible choices for $f(d)$: i) a power-law $f_{\text{pow}}(d) = d^{\alpha}$; and ii) an exponential-law $f_{\text{exp}}(d) = \exp(\alpha d)$. In both cases, the parameter α is obtained by fitting the model to the data in the training set. Because flows span several orders of magnitude, and for the same reasons outlined above for BMS models, we find that training the model on the logarithm of the flows, rather than the flows themselves, leads to more predictive models. Therefore, all results reported here for gravity models correspond to this approach.

Radiation model. We consider the original formulation of model[4], in which flow T_{ij} is modeled as the the total outflow of an origin municipality T_i times the probability of going from i to j. This probability depends on the populations of the origin (m_i) and destination(m_j), as well as the populations of the municipalities within a radius d_{ij} from the municipality at the origin:

$$T_{ij} = T_i\, p_{i\rightarrow j} = T_i\frac{m_i\, m_j}{(m_i + s_{ij})\,(m_i + m_j + s_{ij})}, \tag{7}$$

where $s_{ij} = \sum_{k\neq i,j} m_k(\ \forall\ k\ d_{ik} < d_{ij})$.

Recent works introduce modifications to this model for finite-size systems and in order to avoid border effects[4,32–35] (See Supplementary Fig. S9). However, we find that these models do not outperform the original formulation.

Random forest. We implement a random forest regressor[36] with 1000 estimators. As input data we use a total of 39 features for each origin-destination pair which include: distance, population at origin and destination, and 36 geographical and socio-economic features of the origin and destination areas (see Data and Supplementary Table S5).

Deep Gravity. Taking as a baseline the algorithm developed in ref. 12, we modified the model to predict flows between municipalities rather than between small geographic regions resulting from tessellation. The major difference with the original model is that municipalities are now the smallest geographic unit, allowing us to compare with the models evaluated in this study. In all other aspects (features used, preprocessing of data, an model training) the model remains the same (see Supplementary Text for details, including all parameter values and specific changes in the original code). The modified version of the code can be consulted and downloaded from Symbolic_mobility_BMS/DeepGravity.

Metrics

Common part of commuters (CPC). It is a widely used metric to analyze the performance of mobility models that is defined as

$$\text{CPC} = \frac{2\sum_{ij}\min(T_{ij}, T^{*}_{ij})}{\sum_{ij} T_{ij} + T^{*}_{ij}} \tag{8}$$

where T_{ij} is the predicted value of the flow from i to j and T^{*}_{ij} is the observed flow. The maximum value of the CPC is 1 if there is a complete agreement between the real data and predictions and it decreases to 0 if all predictions for any flow are equal to zero. Note that the CPC is biased toward models that make accurate predictions for large flows, since smaller flows have marginal contributions to the sums. This is especially critical in mobility data where flows can span several orders of magnitude (see Tables 1 and 2).

Absolute error. It measures the distance between the real and predicted data

$$E_{ij} = |T_{ij} - T^{*}_{ij}|\,. \tag{9}$$

Note that absolute error scales with the size of the flows, so that the average absolute error is biased towards the errors of average flows. For this reason, we represent the whole distribution.

Relative error. It measures the difference between real and predicted flows relative to the observed value of the flow:

$$\epsilon_{ij} = \left|\frac{T_{ij} - T^{*}_{ij}}{T^{*}_{ij}}\right|. \tag{10}$$

Note that, while over-predicting flows is penalized by the relative error, under-predicting flows is not, since predicting a zero value for a non-zero flow results in $\epsilon_{ij} = 1$. Because this can bias average relative errors, we plot the whole distribution.

Absolute log-ratio. It measures the difference in the logarithms of predicted and real flows:

$$LR_{ij} = \left|\log\frac{T_{ij}}{T^{*}_{ij}}\right|. \tag{11}$$

Note that for a perfect prediction this metric is equal to zero. Importantly this metric penalizes both over- and under-predictions equally. For instance, a prediction of a flow twice as large $T_{ij} = 2T^{*}_{ij}$ has $LR_{ij} = \log 2$, and a prediction $T_{ij} = T^{*}_{ij}/2$ has $LR_{ij} = \log 2$ as well.

Proportional demographic parity. The goal of this metric is to quantify the fairness of a model when predicting flows between different demographic or socio-economic groups $\{g \in \mathcal{G}\}$. To do so, it quantifies to what extent the errors of the predictions for flows across pairs of groups $\{f_{ij} \equiv (g_i, g_j) \in \mathcal{G}^2 \equiv \mathcal{G} \times \mathcal{G}\}$ are equally distributed for all pairs of (different) groups. Consider that $\bar{l}$ is the median error of all flows, and τ is a percentile window around the median $(0 \leq \tau \leq 100)$. For a pair of flow groups (f_1, f_2), PDP estimate the difference between their error distributions as

$$\mathrm{PDP}_{f_1,f_2} = \left| P\left(\bar{l} - \frac{\tau}{2} \le l \le \bar{l} + \frac{\tau}{2} \middle| f_1\right) - P\left(\bar{l} - \frac{\tau}{2} \le l \le \bar{l} + \frac{\tau}{2} \middle| f_2\right) \right| \qquad (12)$$

where $P(\cdot|f_i)$ is the probability that a prediction of a flow in flow group f_i has an error l such that $\bar{l} - \frac{\tau}{2} \le l \le \bar{l} + \frac{\tau}{2}$.

To get an overall estimate, then PDP uses a weighted average[21]

$$\mathrm{PDP} = \sum_{f,h \in \mathcal{G}^2, f \ne h} w_{f,h} \, \mathrm{PDP}_{f,h} \quad , \qquad (13)$$

where the weight $w_{f,h} = \sum_{k \in \mathcal{G}^2} N_k / (N_f + N_h)$ enhances the relative contribution of small groups of flows. Note that our approach to measure PDP is a generalization of that used in ref. 21, where, instead of percentile windows, the authors consider τ to be a standard deviation around the mean. However, because error distributions are not Gaussian in general (see the explanation for the different error metrics), we use a more general definition applicable to any distribution.

In our analysis, we consider two groups of municipalities: above (rich) and below (poor) the median income per capita (see Data). Therefore we have four different flow groups $\mathcal{G}^2 = \{\text{poor} \rightarrow \text{poor}, \text{rich} \rightarrow \text{poor}, \text{poor} \rightarrow \text{rich}, \text{rich} \rightarrow \text{rich}\}$.

Data availability

All data are available as described in the Methods section. The source dataset is available at GeoDS and is part of the publication Kang et al.[26]. The datasets used to train the models and for the evaluation of the Deep Gravity model are available at Github[37].

Code availability

The code for the BMS is available from https://bitbucket.org/rguimera/machine-scientist. The BMS implementation to obtain mobility models and the Deep Gravity version is available at Github[37].

References

1. Zipf, G. K. The p1 p2/d hypothesis: on the intercity movement of persons. *Am. Sociol. Rev.* **11**, 677–686 (1946).
2. Erlander, S. & Stewart, N. F. The gravity model in transportation analysis: theory and extensions, vol. 3 (VSP, 1990).
3. Guimerà, R., Mossa, S., Turtschi, A. & Amaral, L. A. N. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. USA* **102**, 7794–7799 (2005).
4. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
5. Schläpfer, M. et al. The universal visitation law of human mobility. *Nature* **593**, 522–527 (2021).
6. Yuan, H. & Li, G. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Sci. Eng.* **6**, 63–85 (2021).
7. Haynes, K. E. & Fotheringham, A. S. GraviTY AND SPATIAL INTERACTION MODEls. No. 07 in Wholbk (Regional Research Institute, West Virginia University, 1985).
8. Moro, E., Calacci, D., Dong, X. & Pentland, A. Mobility patterns are associated with experienced income segregation in large us cities. *Nat. Commun.* **12**, 4633 (2021).
9. Balcan, D. et al. Modeling the spatial spread of infectious diseases: the global epidemic and mobility computational model. *J. Comput. Sci.* **1**, 132–145 (2010).
10. Pappalardo, L., Rinzivillo, S. & Simini, F. Human mobility modelling: exploration and preferential return meet the gravity model. *Procedia Comput. Sci.* **83**, 934–939 (2016).
11. Chen, Y. The distance-decay function of geographical gravity model: power law or exponential law? *Chaos Solit Fract.* **77**, 174–189 (2015).
12. Simini, F., Barlacchi, G., Luca, M. & Pappalardo, L. A Deep Gravity model for mobility flows generation. *Nat. Commun.* **12**, 6576 (2021).
13. Pourebrahim, N., Sultana, S., Niakanlahiji, A. & Thill, J.-C. Trip distribution modeling with twitter data. *Comput. Environ. Urban Syst.* **77**, 101354 (2019).
14. Spadon, G., Carvalho, A. C. P. L. Fd, Rodrigues-Jr, J. F. & Alves, L. G. A. Reconstructing commuters network using machine learning and urban indicators. *Sci. Rep.* **9**, 11801 (2019).
15. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
16. Guimerà, R. et al. A Bayesian machine scientist to aid in the solution of challenging scientific problems. *Sci. Adv.* **6**, eaav6971 (2020).
17. Reichardt, I., Pallarès, J., Sales-Pardo, M. & Guimerà, R. Bayesian machine scientist to compare data collapses for the Nikuradse dataset. *Phys. Rev. Lett.* **124**, 084503 (2020).
18. Fajardo-Fontiveros, O. et al. Fundamental limits to learning closed-form mathematical models from data. *Nat. Commun.* **14**, 1043 (2023).
19. Vallès-Català, T., Peixoto, T. P., Sales-Pardo, M. & Guimerà, R. Consistencies and inconsistencies between model selection and link prediction in networks. *Phys. Rev. E* **97**, 062316 (2018).
20. Ho, T. K. Random decision forests. *In*: Proceedings of 3rd international conference on document analysis and recognition. vol. **1**, 278–282 (IEEE, 1995).
21. Liu, Z., Huang, L., Fan, C. & Mostafavi, A. Fairmobi-net: a fairness-aware deep learning model for urban mobility flow generation. https://arxiv.org/abs/2307.11214 (2023).
22. Yabe, T. et al. Enhancing human mobility research with open and standardized datasets. *Nat. Comput. Sci.* **4**, 469–472 (2024).
23. Feng, J. et al. DeepMove: predicting human mobility with attentional recurrent networks. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* 1459–1468, https://doi.org/10.1145/3178876.3186058 (2018).
24. Džeroski, S. & Todorovski, L. (eds.) Computational discovery of scientific knowledge. Lecture Notes in Artificial Intelligence (Springer, 2007).
25. Evans, J. & Rzhetsky, A. Machine science. *Science* **329**, 399–400 (2010).
26. Kang, Y., Gao, S., Liang, Y., Li, M. & Kruse, J. Multiscale dynamic human mobility flow dataset in the U.S. during the COVID-19 epidemic. *Sci. Data* 1–13 (2020).
27. Li, Z., Ning, H., Jing, F. & Lessani, M. N. Understanding the bias of mobile location data across spatial scales and over time: a comprehensive analysis of SafeGraph data in the United States. *PLOS One* **19**, e0294430 (2024).
28. Gallotti, R., Maniscalco, D., Barthelemy, M. & De Domenico, M. Distorted insights from human mobility data. *Commun. Phys.* **7**, 421 (2024).
29. Barreras, F. & Watts, D. J. The exciting potential and daunting challenge of using GPS human-mobility data for epidemic modeling. *Nat. Comput. Sci.* **4**, 398–411 (2024).
30. Grünwald, P. D. The minimum description length principle, vol. 1, (The MIT Press, 2007).
31. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
32. Masucci, A. P., Serras, J., Johansson, A. & Batty, M. Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. *Phys. Rev. E* **88**, 022812 (2013).
33. Yang, Y., Herrera, C., Eagle, N. & González, M. C. Limits of predictability in commuting flows in the absence of data for calibration. *Sci. Rep.* **4**, 5662 (2014).

34. Lenormand, M., Huet, S., Gargiulo, F. & Deffuant, G. A universal model of commuting networks. *PLOS One* **7**, 1–7 (2012).
35. Lenormand, M., Bassolas, A. & Ramasco, J. J. Systematic comparison of trip distribution laws and models. *J. Transp. Geogr.* **51**, 158–169 (2016).
36. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
37. Cabanas-Tirapu, O., Danús, L., Moro, E., Sales-Pardo, M. & Guimerà, R. Github repository: Symbolic_mobility_bms, https://doi.org/10.5281/zenodo.14501438 (2024).

Acknowledgements

We thank L. Pappalardo and M. Luca for help with the Deep Gravity algorithm, and M. Luca for comments and suggestions on the manuscript. This research was supported by projects PID2019-106811GB-C32 (E.M), PID2019-106811GB-C31 and PID2022-142600NB-I00 (M.SP. and R.G.), and FPI grant PRE2020-095552 (O.CT.) from MCIN/AEI/10.13039/501100011033, and by the Government of Catalonia (2021SGR-633) (M.SP. and R.G.). E.M. acknowledges support from the U.S. National Science Foundation under Grants 2218748 and 2420945.

Author contributions

O.C.-T. collected data. O.C.-T. and L.D. wrote code and performed experiments. O.C., L.D., E.M., M.S.-P., and R.G. designed the research, analyzed results, discussed results, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-56495-5.

Correspondence and requests for materials should be addressed to Marta Sales-Pardo or Roger Guimerà.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints