Synthetic demand flow generation using theproximity factor

Ekin Yalvac (eyalvac@ncsu.edu)
North Carolina State University
Michael G Kay
North Carolina State University

Research Article

Keywords: Demand flow generation, origin-destination matrices, spatial interactions

Posted Date: August 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1918195/v1

Synthetic demand flow generation using the proximity factor

Ekin Yalvac[a], Michael G. Kay[a]

July 2022

Affiliations: [a]Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA

Corresponding Author: Ekin Yalvac (ORCID: 0000-0003-0748-8865)
Present Address: 1520 Lilley Ct Apt L1, Raleigh, NC 27606
E-mail Address: eyalvac@ncsu.edu

Abstract

One of the biggest challenges in designing a logistics network is predicting the demand flows between all pairs of points in the network. Until recently, the gravity model was mainly used for estimating the demand flow between points. However, the gravity model uses historical data to estimate values for its multiple parameters and distance between pairs to forecast the demand flow. Distance values close to zero and unprecedented changes in demand flow data create numerical instability for the gravity model's output. Hence, the proximity factor, a single parameter model that uses the relative ordering of pairs instead of distance, was developed. In this paper, we systematically compare the proximity factor and the gravity model. It is shown that the proximity factor is a robust and competitive alternative to the gravity model. According to our analysis, the proximity factor model can replace the gravity model in some applications when no historical data is available to adjust the parameters of the latter.

Keywords: Demand flow generation, origin-destination matrices, spatial interactions

1 Introduction

In a logistics network, the demand flows between all pairs of points in the network are often challenging to determine for existing networks due to a large number of point pairs. In contrast, they can only be predicted when designing a new network. For these reasons, synthetic demand flow generation techniques are used. Gravity-based techniques are most common, where the distance between points is used along with several parameters to estimate the demand flow between points. For an existing network, sampling techniques can be used to estimate multiple parameters. In contrast, when designing a new network, using a single parameter model is preferable so that simple univariate optimization can be used to find the parameter value that results in flows that have an overall desired characteristic like a particular expected average value. Unfortunately, since all gravity methods are based on the distance between pairs of points, distance values at or close to zero result in numerical instability, while extremely large distance outlier pairs are overrepresented in the results. For these reasons, the proximity model was developed as a more robust synthetic demand flow generation technique. Instead of distance, the relative ordering of pairs in terms of their distance is used. This makes it possible to deal with distance values that are zero between pairs of points in the calculation while also not being sensitive to extremes in distance value. While the proximity model has been used in several applications, including the design of a public logistics network [17] and to estimate less-than-truckload (LTL) rates [18], this paper is the first to conduct a systematic comparison of it to the single parameter gravity model.

Designing a logistics network from scratch requires many steps. These steps involve making decisions regarding the number, location, capacity, and technology of distribution centers, warehouses, manufacturing plants, etc [7]. Determining the location of facilities is a crucial part of this process. Given that a set of origin (O) and destination (D) pairs correspond to, for example, potential supply and demand locations of goods in an urban environment, the level of interaction between each O and D pair needs to be estimated. The question addressed in this paper is that, if all that is known is how far, on average, the distance is between all O-D points, how can this be used to estimate the distances between all the pairs of O-D points.

In this paper, we introduce an enhanced version of the previously proposed proximity factor technique [17], where we incorporated the average distance traveled between O-D points. A single parameter technique only requires the proximity factor, ρ, as a parameter. This helps us optimize the locations for distribution centers (DCs) and measure the migration, commuting, or trade flows between locations in real life. Unlike other contemporary models, the proximity factor model does not need previous data. As a result, it is more agile than the multi-parameter gravity model. This yields more accurate predictions for mobility and transport processes [10] [13] [5]. There are also other commodity

and freight flow estimation models using multiple data sources [16] and synthesis of multiple models including the gravity model [11] [12]. The proximity factor differentiates itself from these models as it does not require additional data sources and it does not only calculate freight flow, but any type of demand flow.

The gravity model [9] changed how population movements were predicted when it was first introduced in 1946 [28]. Since then, it has become the most widespread model in this field [6]. Despite this widespread use and popularity in the past, it is far from perfect. It tries to fit the historical data into the model. This leads to oversimplification of flows between O-D pairs, and, in most cases, it fails to grasp actual empirical observations [25] [21] [19]. In addition to these shortcomings, the model needs to estimate multiple parameters. Hence, it is sensitive to fluctuations in data and does not fare well with incomplete data sets [15] [25].

Due to these problems, multiple models have been proposed to replace the gravity model recently. The biggest challenger is the parameter-free radiation model proposed in 2012 [25]. However, the radiation model also lacks the accurate computation of human mobility at the city (or micro) scale [27]. Other factors such as segregation and commercial/residential space distinction are primary drivers of population mobility in a city setting [21].

On the other hand, a simpler single-parameter gravity model has been frequently used in transportation modeling [21]. Since the proximity factor also has a single parameter, we decided that this model could be a good benchmark for the proximity factor's robustness and agility. The optimization of the proximity factor could also be applied to this single-parameter gravity model. As a result, this application puts both techniques on an equal footing for our analysis.

In our analysis, we use four different data sets to assess the proximity factor's accuracy, robustness, and precision addressed to population movements and trade flows. The data sets we use are the Federal Aviation Administration (FAA) Airport Arrival and Departure data [1], US Census commuting data [4], Internal Revenue Service (IRS) migration data [3] and US Department of Transportation Freight Analysis Framework (FAF) data [2].

The paper is structured as follows. Section 2 provides a thorough explanation of the models that were used in our analysis, along with a small numerical example. Section 3 compares both models in terms of their ability to predict demand flows. Section 4 covers the data sets that have been used in our study and shows the output of our comparison. Finally, Section 5 discusses the results and shows leads for further research.

2 Models

We used two models to analyze data sets. Since it is the most well-known model on this topic, we decided to use the outputs of the gravity model to compare the results of the proximity factor method. The proximity factor only uses a single parameter. Thus, we decided to use a single parameter version of the gravity model.

2.1 Proximity Factor

The proximity factor model tracks its origin from genetic algorithms and penalty functions [14]. Transport demand or population migration to and from each O-D pair is estimated by using the percentages of population or a relevant parameter such as the amount of freight that has been transported to each point, together with a proximity factor that controls the degree to which a point is more likely to receive or send flows to nearby points as opposed to points located further away.

There are two main reasons for using the proximity factor. First, to model the impact of distance-related spatial interaction, it provides a single, adjustable parameter. This could be used to alleviate the "edge effect" associated with transport demand that occurs outside the region considered in the analysis. Second, it gives a reference to model the potential effect of the searching and redirection capabilities associated with the operation of a logistics network. For example, an increase in the proximity factor could be used to model the effect of being able to find more items at nearby locations.

Let w_i and w_j be point p_i and p_j's percentage of the total population respectively. Without a proximity factor adjustment, the flow between p_i and p_j is; e.g., the total number of people traveling or tons of freight being transported per certain amount of time, $w^0_{ij} = w_i \times w_j$ and w_{ii} is the demand within the region covered by p_i. Given m points, $p_{[1]}$, $p_{[2]}$,...,$p_{[m]}$, ordered in terms of their increasing great circle distance from p_i, a proximity factor of ρ is used in a normalized geometric distribution [14] as follows:

$$w'_{i[j]}(\rho) = w^0_{i[j]} \frac{(1-\rho)^{(j-1)}}{\frac{1}{m}\sum_{k=1}^{m}(1-\rho)^{(k-1)}} \frac{w_i}{w_i + w_j} \tag{1}$$

$$w_{i[j]}(\rho) = \frac{w'_{i[j]}(\rho)}{\sum_{k=1}^{m}\sum_{l=1}^{m} w'_{kl}(\rho)} \tag{2}$$

$$w_{i[j]_{adj}}(\rho) = w_{i[j]}(\rho) \frac{w_j}{\sum_{k=1}^{m} w_{k[j]}(\rho)} \tag{3}$$

Both $\sum_{i=1}^{m}\sum_{j=1}^{m} w^0_{ij} = 1$ and $\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij} = 1$. The second fraction term

at the right-hand-side of the equation 1 is done to ensure that the resulting $w'_{i[j]}(\rho)$ matrix is symmetrical. The adjustment in the equation 3 is to keep the demand or inflow marginals constant, as demand is tied to the population of that location, and the population does not change. On the other hand, supply may vary as supply is not tied to the population. Without this term, the output of the proximity factor could indicate a drastic change in demand for certain data points. Because population size drives the demand, this addition stabilizes the demand for each data point. Thus, this adjustment can be used where the population is tied to demand and demand does not change with other parameters. However, it is not a fundamental part of the proximity factor calculation.

The proximity factor, ρ, is found via an optimization process, where $\bar{d}$ is average distance function:

$$\bar{d}(\rho) = \sum_{i=1}^{m}\sum_{j=1}^{m} w_{i[j]}(\rho)\frac{w_j}{\sum_{k=1}^{m} w_{k[j]}(\rho)} D_{i[j]} \tag{4}$$

$$\rho^* = \underset{\rho}{\operatorname{argmin}}(\bar{d}(\rho) - \mathrm{d_{avg}}) \tag{5}$$

$$w^*_{i[j]} = w_{i[j]}(\rho^*) \tag{6}$$

where $\mathrm{d_{avg}}$ is the average distance traveled in the existing data set. We multiply distance matrix D with the corresponding likelihood of a node being traveled and take the mean of those values to calculate $\mathrm{d_{avg}}$. The calculation for the cases where there is no historical data is shown in section 3.

2.2 Gravity Model

The gravity model of migration is based on empirical evidence where the commute between locations i and j, with the population of point i being m_i and population of point j being n_j, is proportional to the product of populations of i and j and inversely proportional to the distance function $f(r_{ij})$. The gravity model can be articulated as follows [8]:

$$T_{ij} = \frac{m_i^{\alpha} n_j^{\beta}}{f(r_{ij})} \tag{7}$$

α, β, and $f(r_{ij})$ are determined through multiple regression to fit the data. $f(r_{ij})$ can take the form r_{ij}^{γ} or $e^{r_{ij}/\kappa}$. γ and κ can also be found via multiple regression to fit the empirical data.

2.2.1 Single Parameter Gravity Model

The proximity factor has only a single parameter ρ. We used a single-parameter gravity model to make the gravity model comparable to the proximity factor. This model only uses the parameter as the exponent of the denominator, and

it is more widely used in transportation modeling:

$$T_{ij} = \frac{m_i n_j}{r_{ij}^{\gamma}} \tag{8}$$

Similar to the proximity factor, γ is found via an optimization process where $\bar{d}$ is average distance function:

$$\bar{d}(\gamma) = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{T_{ij}(\gamma)}{\sum_{k=1}^{m} \sum_{l=1}^{m} T_{kl}(\gamma)} D_{ij} \tag{9}$$

$$\gamma^* = \underset{\gamma}{\operatorname{argmin}} (\bar{d}(\gamma) - \mathrm{d_{avg}}) \tag{10}$$

$$T_{ij}^* = T_{ij}(\gamma^*) \tag{11}$$

Since the output of the gravity model is symmetrical, the adjustment we made in the proximity factor section is not necessary for the gravity model.

2.3 Numerical Example

To show the difference between the proximity factor and the gravity model outputs, we decided to apply them in a small-town setting. The best candidate for us was Spencer–Spirit Lake, IA Combined Statistical Area (CSA). This is the smallest CSA out of 172, and when we eliminate the low-density parts of the area, there are 6 data points left for us to analyze. As a result, Spencer–Spirit Lake is the perfect candidate for us to show the difference between the two models without making the outputs too complicated.
The distance matrix is calculated by great circle distance calculation:

$$D = \begin{bmatrix} 0.000 & 1.015 & 1.576 & 2.190 & 1.538 & 1.428 \\ 1.015 & 0.000 & 0.655 & 1.175 & 0.525 & 0.671 \\ 1.576 & 0.655 & 0.000 & 0.798 & 0.476 & 1.055 \\ 2.190 & 1.175 & 0.798 & 0.000 & 0.655 & 1.080 \\ 1.538 & 0.525 & 0.476 & 0.655 & 0.000 & 0.606 \\ 1.428 & 0.671 & 1.055 & 1.080 & 0.606 & 0.000 \end{bmatrix}$$

The population data for every single data point is

$$Pop = \begin{bmatrix} 1082 \\ 767 \\ 763 \\ 621 \\ 718 \\ 787 \end{bmatrix}$$

The average distance (d_{avg}) for this setting is calculated to be 0.4877 miles. We took the geometric mean of a lower bound and an upper bound value of two different average distance calculations to reach this value. We assumed that a grocery store would serve roughly 1000 people in this town. The total

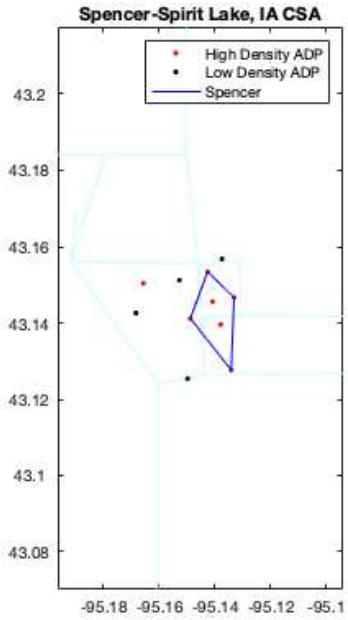


Figure 1: Spencer–Spirit Lake, IA CSA data points

population of the city of Spencer, excluding the low-density data points, is 4738. This would roughly translate to 4 grocery stores in this town. The total area of the 6 data points is 4.96 square miles. This yields 1.24 square miles per grocery store. In the lower bound calculation, we assumed that people would travel two-thirds of the area diameter on average to reach a grocery store, which corresponds to the average distance to the center of a circle, assuming a constant population density. Hence, the calculation for the lower bound is

$$d_0^{LB} = \frac{2r}{3} \tag{12}$$

$$r = \frac{3d_0^{LB}}{2} \tag{13}$$

$$a = \pi \left(\frac{3d_0^{LB}}{2} \right)^2 \tag{14}$$

$$d_0^{LB} \sim 0.376\sqrt{a} \tag{15}$$

The upper bound calculation is found by the formula [22]

$$d_0^{UB} = \frac{32}{15} \frac{\sqrt{a}}{\pi \Gamma(5/2)} \sim 0.51\sqrt{a} \tag{16}$$

When we used these formulae, we get d_0^{LB} as 0.419 and d_0^{UB} as 0.568. Thus, the geometric mean is provided as an average distance of 0.488.

For the proximity factor calculation, we needed a population weight vector, so we used the population vector to calculate weights. When we divide the population vector by the total population, we get a vector w_i of

$$w_i = \begin{bmatrix} 0.228 \\ 0.162 \\ 0.161 \\ 0.131 \\ 0.152 \\ 0.166 \end{bmatrix}$$

Using the calculations stated in section 2.1 (Equations 4 and 5), we found ρ^* to be 0.396. The w_{ij}^0 matrix is found by the matrix multiplication of the w_i vector and transpose of it. After this process, we get

$$w_{ij}^0 = \begin{bmatrix} 0.052 & 0.037 & 0.037 & 0.030 & 0.035 & 0.038 \\ 0.037 & 0.026 & 0.026 & 0.021 & 0.025 & 0.027 \\ 0.037 & 0.026 & 0.026 & 0.021 & 0.024 & 0.027 \\ 0.030 & 0.021 & 0.021 & 0.017 & 0.020 & 0.022 \\ 0.035 & 0.025 & 0.024 & 0.020 & 0.023 & 0.025 \\ 0.038 & 0.027 & 0.027 & 0.022 & 0.025 & 0.028 \end{bmatrix}$$

After equation 1, we find the matrix $w'_{i[j]}$.

$$w'_{i[j]}(0.396) = \begin{bmatrix} 2.497 & 0.820 & 0.255 & 0.201 & 0.340 & 0.500 \\ 0.820 & 2.497 & 0.911 & 0.273 & 1.200 & 0.728 \\ 0.255 & 0.911 & 2.497 & 0.749 & 1.508 & 0.440 \\ 0.201 & 0.273 & 0.749 & 2.497 & 0.963 & 0.454 \\ 0.340 & 1.200 & 1.508 & 0.963 & 2.497 & 1.007 \\ 0.500 & 0.728 & 0.440 & 0.454 & 1.007 & 2.497 \end{bmatrix}$$

Once we find the matrix $w'_{i[j]}$, we normalize this matrix to get the percentage of total flows between points i and j.

$$w_{i[j]}(0.396) = \begin{bmatrix} 0.133 & 0.031 & 0.010 & 0.006 & 0.012 & 0.019 \\ 0.031 & 0.067 & 0.024 & 0.006 & 0.030 & 0.020 \\ 0.010 & 0.024 & 0.066 & 0.016 & 0.038 & 0.012 \\ 0.006 & 0.006 & 0.016 & 0.044 & 0.020 & 0.010 \\ 0.012 & 0.030 & 0.038 & 0.020 & 0.059 & 0.026 \\ 0.019 & 0.020 & 0.012 & 0.010 & 0.026 & 0.071 \end{bmatrix}$$

Finally, we apply the inbound demand adjustment demonstrated in the equation 3 to ensure the demand information stays the same.

$$w_{i[j]_{adj}}(0.396) = \begin{bmatrix} 0.144 & 0.028 & 0.009 & 0.008 & 0.010 & 0.020 \\ 0.034 & 0.061 & 0.024 & 0.008 & 0.025 & 0.021 \\ 0.010 & 0.022 & 0.064 & 0.021 & 0.031 & 0.013 \\ 0.007 & 0.005 & 0.016 & 0.056 & 0.016 & 0.011 \\ 0.013 & 0.027 & 0.037 & 0.025 & 0.048 & 0.027 \\ 0.021 & 0.018 & 0.012 & 0.013 & 0.021 & 0.074 \end{bmatrix}$$

As a result, we conclude that the output of proximity factor is the matrix $w_{i[j]_{adj}}(0.396)$. We see the effect of inbound adjustment when we take the sum of all columns.

$$\sum_{i=1}^{m} w_{i[j]_{adj}}(0.396) = \begin{bmatrix} 0.228 & 0.162 & 0.161 & 0.131 & 0.152 & 0.166 \end{bmatrix}$$

We can also get the same result by taking the sum of all columns of w_{ij}^0 matrix.

$$\sum_{i=1}^{m} w_{ij}^0 = \begin{bmatrix} 0.228 & 0.162 & 0.161 & 0.131 & 0.152 & 0.166 \end{bmatrix}$$

This indicates that we kept the demand constant in the proximity factor calculation, whereas supply is versatile.

w	wij^0		$w_{ij_{noadj}}$		$w_{ij_{adj}}$	
	w_i^0	w_j^0	$w_{i_{noadj}}$	$w_{j_{noadj}}$	$w_{i_{adj}}$	$w_{j_{adj}}$
	0.228	0.228	0.212	0.212	0.228	0.220
	0.162	0.162	0.178	0.178	0.162	0.171
	0.161	0.161	0.166	0.166	0.161	0.161
	0.131	0.131	0.102	0.102	0.131	0.111
	0.152	0.152	0.184	0.184	0.152	0.178
	0.166	0.166	0.158	0.158	0.166	0.159

Table 1: Sum of i and j dimensions with and without inbound adjustment

We use the same parameters used in the proximity factor calculations, such as average distance and population data, for the gravity model calculations. γ^* is found via a similar process that has been used for ρ^*. This optimization process yields a γ^* value of 16.998. Applying this value to the equation 8 yields a T_{ij} matrix. We normalize this matrix to make it comparable to the output of the

proximity factor. Hence, we get the matrix $T_{ij_{per}}$

$$T_{ij_{per}} = \begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.002 & 0.000 & 0.079 & 0.001 \\ 0.000 & 0.002 & 0.000 & 0.000 & 0.409 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.001 & 0.000 \\ 0.000 & 0.079 & 0.409 & 0.001 & 0.000 & 0.007 \\ 0.000 & 0.001 & 0.000 & 0.000 & 0.007 & 0.000 \end{bmatrix}$$

These two models try to measure the same phenomenon, the percentage of flows that occurs between these six points. We expect them to be somewhat correlated as they used the same data to predict the exact relationship between the points. However, the correlation coefficient between these two models is merely 0.052. This essentially means that the outputs of these models are not correlated. A question arises from this result. Which model is more reliable in predicting flows between locations? To answer this, we decided to see how these models fare with the real-life data in section 4.

3 Model Comparison

We reran these algorithms twice in a row using different parts of the same data set to compare the robustness between the proximity factor and the gravity model. For the sake of simplicity, we chose Gainesville, FL, in most of our calculations. This area is isolated from other population centers and small enough for us to run both models rapidly. Gainesville–Lake City, FL Combined Statistical Area consists of 93 high-density aggregate data points with a population of 372,607 people. We ran the entire data set in our first iteration to predict the parameters ρ and γ. We then divided data points into test and train data sets where approximately 70% of the data is allocated for train, and the rest is allocated for test data set unless otherwise specified. We then found parameters again using the train data set and applied the algorithms again. This constituted our second iteration. Finally, we compared the estimated proximity factor/gravity results to the first iteration's test data to the second iteration's outputs to decide which algorithm is more reliable. We used this process in different settings to ensure our results were unassailable.

Initially, we applied these algorithms to three different-sized cities: Gainesville, FL, being small-sized; Raleigh, NC, being medium-sized; and Atlanta, GA, being large-sized. Table 2 shows the results of these cities. In these applications, we kept the train set ratio at 70%. The top two rows of Table 2 show us the root mean square error (RMSE) of the entire data set for the proximity factor and gravity models. The middle two rows show the RMSE of the diagonal outputs of the proximity factor and gravity models (corresponding to the local demand). The bottom two rows show the RMSE of non-diagonal elements. All error values are relative and shown in percentages. It can be easily seen that the proximity

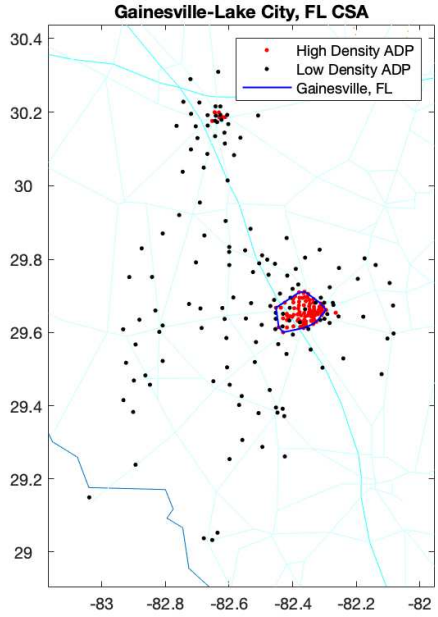


Figure 2: Gainesville–Lake City, FL CSA data points

RMSE (in %)	Gainesville, FL	Raleigh, NC	Atlanta, GA	LTL
P	7.68	0.96	6.26	26.77
G	9.59	14.05	10.81	174.93
P_{diag}	1.49	0.11	0.28	–
G_{diag}	6.42	8.46	5.41	–
$P_{nondiag}$	7.72	0.96	6.26	–
$G_{nondiag}$	9.62	14.06	10.82	–

Table 2: Relative root mean square error (RMSE) results for different cities' gravity and the proximity factor models.

factor algorithm has smaller RMSE in all applications. This indicates that the difference between the two iterations for the proximity factor is smaller, making it more reliable than the gravity model. We also checked the weight of the diagonal of output matrices since the diagonal shows the local demand where the distance is the smallest. We found that the weight of diagonal for the proximity factor ranged from 2% to 12%, and for the gravity model, it ranged from 34% to 37%. This considerable difference shows that the gravity model is biased towards the local demand, whereas the proximity factor distributes the demand evenly in the region. Also, we compared the proximity factor and gravity model

for the nationwide less-than-truckload (LTL) shipments. In this case, we used 3-digit ZIP codes and a fixed ρ value of 2.57. The average distance value of 752 miles is used which is the average LTL shipment distance throughout the US [26]. In this case, we only show the overall RMSE, as there would not be any significant LTL shipment for the local demand. The proximity factor fares better compared to the gravity model in this case as well. For the remainder of our comparison, we only used Gainesville, FL area.

	d_{avg}	$90\% \times d_0^{UB}$	$110\% \times d_0^{LB}$
$RMSE_P$	3.035×10^{-6}	7.965×10^{-7}	6.179×10^{-5}
$RMSE_G$	2.472×10^{-5}	9.951×10^{-6}	2.626×10^{-5}

Table 3: Root mean square error results for the gravity and the proximity factor models for upper and lower bound values of average distance in Gainesville, FL area.

In the next set, we tried to see the extremes of the average distance. The calculation of it is given in the first numerical example. In this example we used 90% of d_0^{UB} and 110% of d_0^{LB} to check for the extremes. Except for the lower-bound extreme, the proximity factor is faring better. The gravity model works better with lower average distance values since it heavily favors shorter distance demand over longer distance demand. Hence, it is better at a low average distance. However, the proximity factor model is within the same order of magnitude, making it compatible with the gravity model, if not better.

In another setting, we tried different percentages of train sets. In the second iteration, we used the train set to obtain parameters for both models. Using different sizes of train sets may affect the robustness of the gravity model or the proximity factor algorithms. In Table 4, we see an increase in RMSE with lower data points in a train set. However, the proximity factor always gives us a lower value for an RMSE showing a more robust performance.

Until this point, we used the aggregate distance as the main form of distance calculation. In our last setting, we used the great-circle distance in our calculations. This came with its shortfalls. For local demand, the great-circle distance calculation comes up with zero value. This is not a problem for the proximity factor algorithm, but for the gravity model, it yields infinity as the distance is the denominator in the model. To avoid this, we simply assumed zero value for the diagonal of the gravity model output. This gave an advantage to the gravity model in the RMSE calculation. However, the proximity factor still fares better in this setting with $RMSE_P$ of 6.314×10^{-6} compared to the gravity's $RMSE_G$ of 1.581×10^{-5}. These results indicate that the proximity factor is as reliable as the gravity model.

For the final part of our analysis of this example, we compared the error differ-

Train set %	$RMSE_P$	$RMSE_G$
10	9.328×10^{-5}	2.423×10^{-4}
20	4.523×10^{-5}	1.543×10^{-4}
30	3.922×10^{-5}	1.429×10^{-4}
40	1.536×10^{-5}	9.660×10^{-5}
50	7.436×10^{-6}	5.241×10^{-5}
60	3.751×10^{-6}	4.216×10^{-5}
70	3.035×10^{-6}	2.472×10^{-5}
80	1.835×10^{-6}	1.959×10^{-5}
90	4.068×10^{-6}	1.716×10^{-5}

Table 4: Root mean square error results for the different training sets' gravity and the proximity factor models.

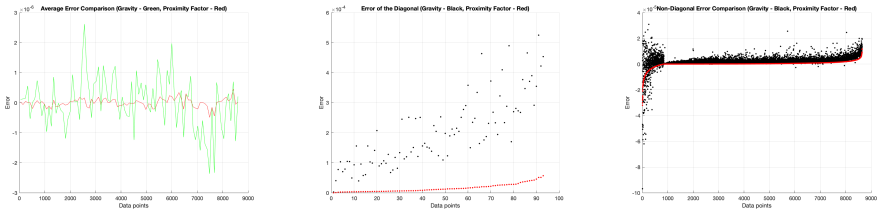


Figure 3: Error Comparison of Gainesville run: Average error comparison in the left, Diagonal points error comparison in the middle, Non-diagonal error comparison in the right

ence between the two iterations of the gravity and the proximity factor models. In this case, we only used the Gainesville example with a 70% training set, aggregate distance calculation, and average distance calculation outlined in example 1. The results can be seen in Figure 3. The figure on the left shows us the average error per demand point. We take the average of every 93 points in the output matrix. It can be easily seen that the error variance of the proximity factor is much lower than the error variance of the gravity model. The figure in the middle shows us the error in the diagonal points where we calculate the local demand. The error is much more significant for those points as we expect the greatest demand would occur at a point where the distance is the smallest. In this graph, we arranged the points according to their ascending error in the proximity factor calculation. The proximity factor error seems to be always smaller than the gravity model's error. This also shows us that the proximity factor is more robust than the gravity model. The figure on the right shows us the error in the non-diagonal points. The demand for these points is smaller than the diagonal as goods need to travel a certain distance to reach their destination. We made a similar ascending error adjustment as we did for the figure in the middle. Still, we can see that the error for the proximity factor is smaller than the gravity model except for a handful of points. These examples and the

analyses show us that the proximity factor is as reliable or robust as the gravity model, if not better. Since we showed that both models are compatible, we decided to proceed with the comparison of these two models with real-life data in section 4.

4 Data Analysis

In this section, we tested the models defined in equations 2 and 8 against empirical data. We used the US commuting data from US Census [4], IRS migration data [3], FAA airport arrival and departure data [1], and US Department of Transportation's Freight Analysis Framework (FAF) data [2].

The US Census commuting data has more short-distance flows than long-distance ones. Therefore, it focuses more on cities than the entire US. This data set contains more than 137 000 rows. Due to the immense amount of data, our computers could not process the whole nation. Thus, we decided to focus on New York City (NYC) and the locations to which people from NYC travel. With its vast population, NYC provided a good data set of flows focusing on short distances such as tri-state area commute (NY, NJ, and CT) and long distances such as NYC to New Mexico. The only caveat is the data granulation, where only county-level flows could be analyzed.

The IRS migration data give information about people who filed their taxes from different addresses in consecutive years. IRS considers these people moved from one location to another. This data has a similar granulation issue as the Census data. It only provides migration between counties. As a result, we could not analyze the inter-city migration unless a city lies in multiple counties. To overcome this obstacle, we focused on analyzing migration patterns for people moving in and out of Atlanta. Atlanta lies in multiple counties, and it is large enough to attract people from the entire country. However, migration has other parameters, such as economy, policies, worldwide events, et cetera, that neither the proximity factor nor the gravity model can pick up in their analyses of this data set.

FAA's US Airport departure-arrival data would give us information about how much some of the airports were used in the United States. This data is not considered good enough for the proximity factor and the gravity model for several reasons. First, short distances were not covered as people preferred air travel for longer distances. Second, small cities are misrepresented as large cities that have larger flows. The smaller a city gets, the exponentially fewer flows occur for that city. Third, flights between cities have different parameters besides population and distance. Fourth, some large cities have multiple airports, which skews the data as we consider both airports serve the same population. Fifth, the data does not provide layover information, so we are not sure how many people traveled to a particular city or had a layover at that airport. Hence, we con-

cluded that this is not a good data set and did not proceed with further analysis.

FAF provides us with how much freight (or commercial goods) is transported between FAF zones. There are 132 FAF zones throughout the US. As a result, origin and destination points are not clearly defined, and the granularity of the data is low compared to other data sets we analyzed. On the other hand, this is the only data set that gives us insight into the commodity flow in the United States.

4.1 Census Data

In this section, we test the gravity model defined in equation 8 and the proximity factor defined in equation 2 against the empirical commute data for the United States. The data was collected by the Census Bureau and had the commute information of where the people living in a specific county travel for work. The data has county pairs that indicate the origin and destination county FIPS codes. Using Matlog's [23] *uscounty* database, we used those FIPS codes to find the location of the corresponding county.

This database has over 139 000 origin-destination pairs, and the US has more than 3000 counties. Using this entire database and creating a distance matrix of 3000 by 3000 was not feasible for us and our computers. Hence, we decided to focus on the commuting information of people living in New York City. People are traveling to and from New York City from the entire country. Thus, it provides sufficient diversity of locations being commuted. Furthermore, the tri-state region that encompasses New York City has one of the highest population densities in the US. This type of characteristic also provided a wide range in the number of people traveling between counties.

According to this data, people living in New York County travel to 154 other counties in the nation. Including New York County, we created a great circle distance matrix of 155 by 155. To ensure that all these distance pairs have relevant flow information, we found all other people commuting between these 155 counties. We created another 155 by 155 matrix consisting of all the flows between these counties.

	Gravity	Proximity Factor
Airport	0.0531	0.0754
Census	0.0021	0.9595
IRS	0.5550	0.1628
FAF	0.0093	0.5919

Table 5: R^2 results for the gravity and the proximity factor models for different data sets

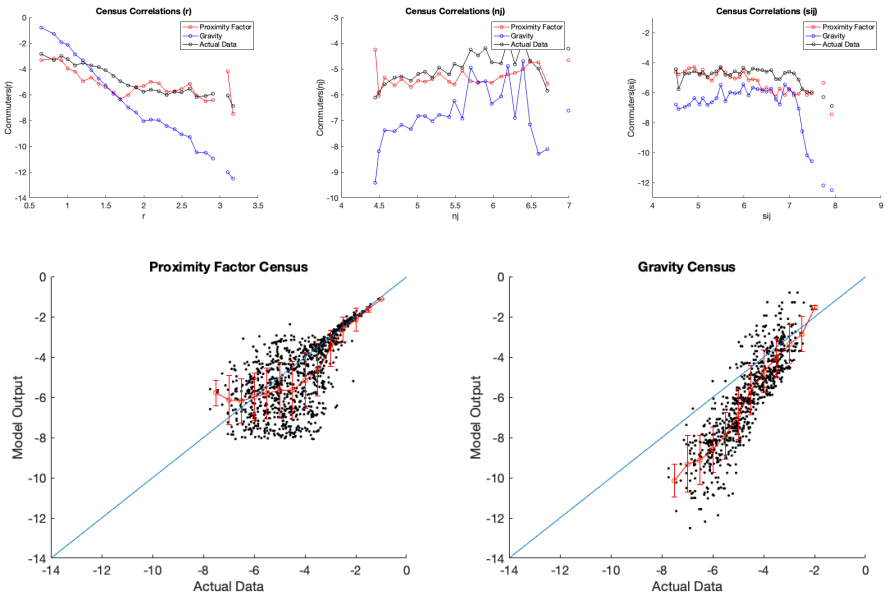


Figure 4: Analysis results for Census data. Top panel Census commuter flow 2011-15. Parameters for the single parameter gravity model: γ=4.656

In Table 5, we can see that the proximity factor is superior to the gravity for the Census data set. The proximity factor output could explain 95% of the variation in data, whereas the gravity is virtually at 0%. Also, the mean square error for the proximity factor ($RMSE_P$) is 1.723×10^{-4} whereas for the gravity model ($RMSE_G$) is 1.9×10^{-3} indicating the proximity factor has a significantly lower error. One of the main reasons is that the gravity model cannot pick up the commute inside a county. The model predicts that anything that happens within a county is zero since D_{ii} is zero. This omits quite a substantial amount of data as most people in the US commute within their county. The proximity factor does consider that factor. It puts a lot of emphasis on intra-county commutes. To test this hypothesis, we developed the following aggregate distance ($\mathrm{d_{agg}}$) estimate:

$$\mathrm{d_{agg}} = \begin{cases} \dfrac{2R}{3} + \dfrac{D}{48} + \dfrac{9D^2}{20R}, & \text{if } D < R \\ \\ D + \dfrac{3R^2}{23D}, & \text{otherwise.} \end{cases}$$

The estimate determines the average distance from a point located at a distance D from the center of a circular region with a radius R to all of the points in the region, assuming the points are uniformly distributed in the region. Since there is no simple analytical formula to estimate this average distance, except for the case of the point being at the center of the region, in which case $D = 0$ and the

average distance is two-thirds of the radius, the estimate represents a regression on a uniformly distributed random sample of points in the region. In our case, the area of a county is approximated as a circular region. This calculation avoids the zero value for D_{ii}. When we used this for distance calculation in our code, the gravity model's R^2 performance improved from 0.0021 to 0.2126. The results of this analysis can be seen in Table 6.

Other factors play into the proximity factor's superiority over gravity. The gravity model practically imitates the data. It requires a multiple or linear regression to calculate its parameters. The proximity factor is independent of these types of factors. All it requires is the distance information among points and population percentages of points in space. This type of independence makes it robust compared to the gravity model. This can be seen in correlations in Fig. 4. Here, we show the correlations of the commuting flows with three sensitive quantities: the distance r in the left panel, the destination population n_j, and the population in the circle centered on the origin population, with radius r, s_{ij}. All the plots are in a log-log scale, so the correlations are in the form of power laws.

The proximity factor performs well in reproducing these correlations, while the gravity model fails to do so in this data set. Considering the same number of parameters required for the proximity factor and the gravity model, the former performs better than the latter.

In the bottom panel of Fig. 2, we show the analysis of the flows of commuters from New York City. Here, we show a scatter plot comparison between the actual data (x axis) and the model's output (y axis). Moreover, we show the mean in red circles, standard deviation bars, and the x=y line to show where the model meets the actual data. All the plots in this panel are also on a log-log scale. Here we can see that the proximity factor fares considerably well with larger flow estimation. When the number of people traveling decreases, the precision also decreases but considering this is a log-scaled graph, the error is less drastic than it seems. However, the gravity model tells us a different story. The accuracy of the model appears to be off. Even though it gets better with more significant flows, it is not as good as the proximity factor, and it outright underperforms with smaller flows.

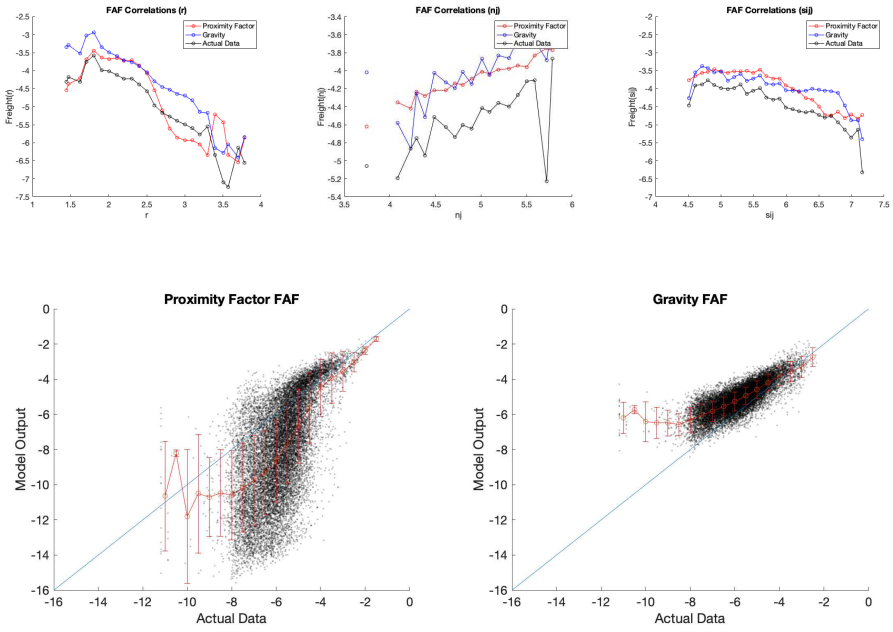
4.2 FAF Data

Here, we test the models above on the FAF data set. The data was collected by the US Department of Transportation and had the freight transportation information that had occurred in the country between 2012 and 2017 [2]. The data has 132 FAF Zones. Zones consist of major metropolitan areas for high population density regions and state or rest of the state for low population density regions. There are no multi-state zones in this data set.

Besides the O-D pairs, the data also includes the amount of freight traveled between these points as tons and ton-miles. However, there were some gaps in our distance matrix as some zones did not have commodity transfers. These gaps were filled with the data acquired from Google Maps. We calculated the distances between O-D pairs by dividing ton-miles by tons. Since the number of FAF Zones was small enough to calculate the entire country, we did not focus on a particular region as we did for other data sets. We also used the data from the 2007–12 databases for comparison purposes.

	Gravity	Proximity Factor
Census	0.2126	0.9595
IRS	0.0936	0.1628
FAF	0.5109	0.4971

Table 6: R^2 results for the gravity and the proximity factor models for different data sets using d_{agg} for the distance calculation



Our analysis concludes that the proximity factor is superior to the gravity model (Table 5). R^2 values indicate that the proximity factor could explain 59% of the variation in data, whereas the gravity model is in the range of 0%. In addition to that $RMSE_P$ is 3.374×10^{-4} and $RMSE_G$ is 6.969×10^{-4} showing less error for the proximity factor. The main reason for this discrepancy might be the

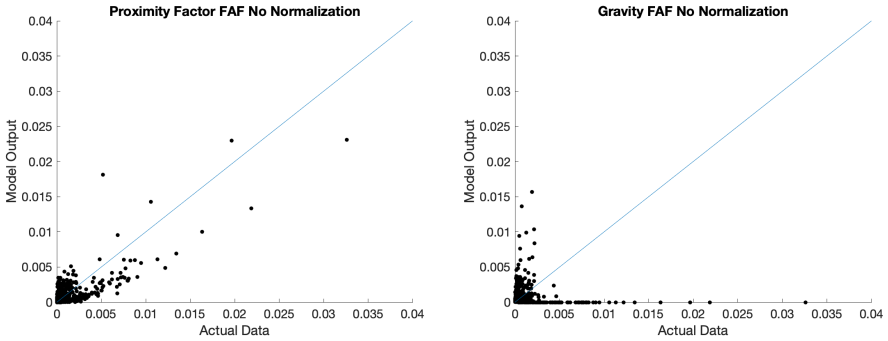


Figure 5: Analysis results for FAF data. Top panel FAF5 2012-17. Parameters for the single parameter gravity model: γ=2.5346

distance calculation. Distance is calculated using the data provided (ton-miles divided by tons), but we could not find a satisfactory explanation of how it is calculated. Hence, we decided to calculate the distances between zones on our own. We found a list with county information for all the FAF zones. Using their FIPS codes and Matlog's *uscounty* database, we calculated the geographical centers of all zones. Next, we used the great-circle distances to calculate distances between O-D pairs. This method neither changed the R^2 results for the proximity factor nor the gravity model as the former was around 0.54 and the latter was still at 0.00.

Great-circle distance calculation proved to be more problematic than our first distance calculation. In our initial analysis, where we used tons and ton-miles information, we considered intra-zonal freight transportation. Nonetheless, the great-circle calculation gave us zero distance for intra-zonal transport. As we mentioned earlier, this is not a problem for the proximity factor, but it affects the gravity model calculations. To make it an even playing field for both methods, we again used the aggregate distance function, d_{agg}. Coincidentally, this analysis gave an R^2 value of 0.51 for the gravity. The proximity factor R^2 value did not change much as it stayed around 0.50. This drastic change showed us how the proximity factor is vigorous compared to the gravity model and how the gravity model is sensitive to the distance calculation between O-D pairs.

When we look at the correlations in the top panel of Figure 5, the proximity factor is faring better with distance r than the gravity model. However, the gravity model gets better with increasing r values. The destination population, in this case, ton value, n_j, correlations show us that both the proximity factor and the gravity model are not doing well. s_{ij} correlations also show no indication of the superiority of one model over the other. Some parts are where the proximity is doing better, and somewhere the gravity is doing better. As the Census data correlations, all the plots are in log-log scale, so the correlations

are in the form of power laws.

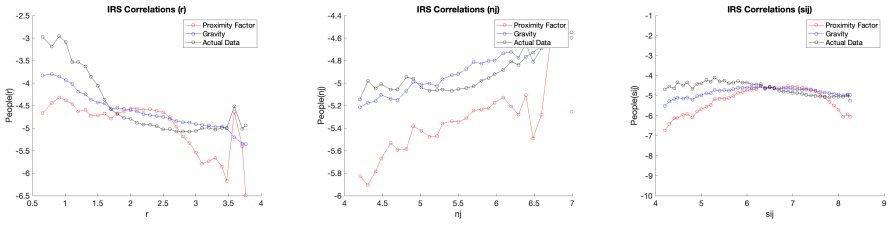
In the middle panel of Fig. 5, we analyze the freight transportation routes between the zones. All the plots in this panel are also in the log-log scale, and the necessary explanation was made in the Census data section for these graphs. Here we can see that the proximity factor is more accurate than the gravity model, but the precision is lower for the proximity factor. Both models get better with higher ton values, but the proximity factor's accuracy increases substantially compared to the gravity model. However, these graphs can be misleading. Hence, we added the actual scatter plots in the bottom panel; the proximity factor's ability to predict the actual data can be successfully seen there. The noise at the bottom left section caused the log-log scale to clutter unnecessarily.

4.3 IRS Data

Here, we test the proximity factor and the gravity models on IRS migration data. The way the data was collected is very similar to the Census data. Consequently, we used the same methodology to calculate the distances among counties.

Similar to Census data, this database has over 113 000 origin-destination pairs. As a result, we decided to focus on Fulton County, where downtown Atlanta is located. Atlanta is a huge city with suburbs laying over 13 counties. This characteristic allowed us to look at the migration occurring both within the city and in other parts of the country.

According to this data, people living in Fulton County migrated to 317 counties around the country. Including the starting county itself, we created a distance matrix of 318 by 318. Then, we found all the other people migrating between these 318 counties. We created another 318 by 318 matrix consisting of all the migration between these counties.



By glancing at the correlations graphs in Figure 6, it can be easily seen that the gravity model is faring better than the proximity factor, especially with an increasing population. The proximity factor seemed unreliable with increasing values as it diverges from the actual data correlations. These graphs are also

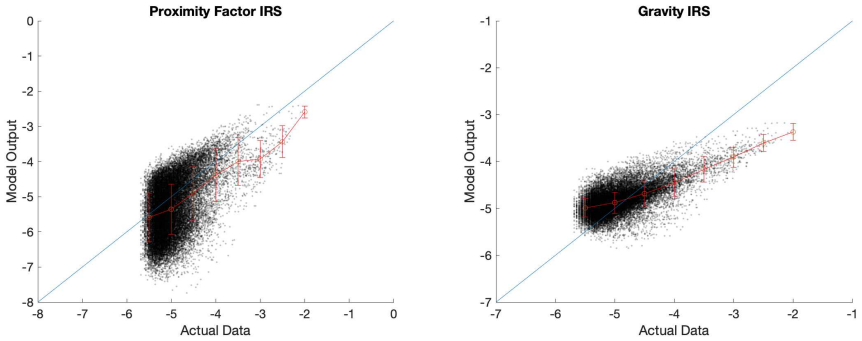


Figure 6: Analysis results for IRS data. Top panel IRS migration data 2010-11. Parameters for the single parameter gravity model: γ=0.85

on a log-log scale. Hence, the correlations are in the form of power laws. In the bottom panel of Fig. 6, we show the scatter plots for the proximity factor and the gravity model compared to the actual data. The gravity model is better aligned with the actual data as we see more points closer to the x-y line. When we look at the proximity factor graph, we can see that the upper end of the scatter plot is not well aligned with the x-y line as it did with other data sets.

In Table 5, we can see that the gravity model is superior to the proximity factor for the IRS data set. The gravity model output can explain around 55% of the variation in the data, whereas the proximity factor can only do 16%. Also, $RMSE_P$ is 9.997×10^{-5} and $RMSE_G$ is 7.042×10^{-5} indicating less error for the gravity model. The main reasons for this discrepancy are explained above. Inter or intracity migration has additional factors besides population and distance. The gravity model and the proximity factor only consider population and distance as a factor. The COVID-19 pandemic showed us that migration patterns are tough to predict as people move to suburbs or smaller cities to work from home [20]. Since we tested aggregate distance (d_{agg}) in other data sets, we wanted to see how would the gravity model fare with it. The R^2 value did not change for the proximity factor. However, the gravity model's R^2 value went down from 55% to 9%. This is another indicator that the proximity factor is more robust than the gravity model. The difference in distance calculation also puts in doubt the reliability of the gravity model's output data.

5 Conclusion

There are multiple aspects to the synthetic demand generation problem. Various models have been proposed recently, but the gravity model is still considered the best option to handle this problem. The proximity factor's independence from the metric distance and requiring fewer parameters than conventional models, such as the gravity, is a significant and desired change.

In this paper, we address the compatibility and reliability of the proximity factor against the gravity model for different data sets.

The first thing we noticed was that the dynamics of the data set play a considerable role for most models in computing the flows. If there are flow patterns, models do not fare well with that particular data set.

There are also some structural problems associated with the gravity model. It changes drastically with the change in distances and cannot calculate the same point interactions where distance is zero. The gravity model's use of past data is another problem. It tries to imitate past data but cannot absorb the abrupt changes happening at the moment. The proximity factor does not require a past data set to use in applications. This gives a considerable advantage. The robustness of the proximity factor is shown in the data analysis.

We could not utilize the airline passenger origin-destination data as FAA only provides airport arrival and departure data. Airline companies do not disclose passenger start-end point data. As a result, FAA data was not very useful for the analysis. This can also be seen in R^2 results for Airport data. If we could access passenger start-finish data, it would be interesting to see the results of comparing the gravity model and the proximity factor in the airline industry.

Despite its satisfactory performance, the proximity factor can be further improved. Its flow ranking algorithm can be improved to reflect other environmental changes, such as changing it from order ranking to distance ranking. In this regard, a new universal visitation law has been proposed recently [24]. The inverse correlation between the multiplication of distance and population density can be integrated into the proximity factor. However, the universal visitation law considers the probability of visitation for individuals, whereas the proximity factor calculates the likelihood of demand flows. As a result, we believe this integration would be the out of scope of this paper.

Another further improvement could be the investigation of the universality of the proximity factor. Analyzing its implementation in different settings and environments would prove whether it is universal.

Acknowledgments

We want to thank Ganeshan Subramanian for kick-starting the proximity factor analysis.

Statements and Declarations

Funding

The authors have no relevant financial or non-financial interests to disclose.

Competing Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Availability of Data and Material

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Authors' Contribution

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Ekin Yalvac, and Michael G. Kay. The first draft of the manuscript was written by Ekin Yalvac and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] Bureau of Transportation monthly transportation statistics. https://data.bts.gov/stories/s/m9eb-yevh. Accessed: 2020-11-30.

[2] Department of Transportation freight analysis framework. https://faf.ornl.gov/faf5/. Accessed: 2021-03-18.

[3] Internal Revenue Service soi tax stats - migration data. https://www.irs.gov/statistics/soi-tax-stats-migration-data. Accessed: 2021-02-23.

[4] US Census Bureau commuting data. https://www.census.gov/topics/employment/commuting/data.html. Accessed: 2020-12-31.

[5] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[6] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

[7] Jean-François Cordeau, Federico Pasin, and Marius M Solomon. An integrated model for logistics network design. *Annals of operations research*, 144(1):59–82, 2006.

[8] Juan de Dios Ortúzar and Luis G Willumsen. *Modelling transport*. John wiley & sons, 2011.

[9] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[10] Weisi Guo, Bogdan Toader, Roxana Feier, Guillem Mosquera, Fabian Ying, Se-Wook Oh, Matthew Price-Williams, and Armin Krupp. Global air transport complex network: multi-scale analysis. *SN Applied Sciences*, 1(7):1–14, 2019.

[11] Jose Holguin-Veras and Gopal R Patil. Integrated origin–destination synthesis model for freight with commodity-based and empty trip models. *Transportation Research Record*, 2008(1):60–66, 2007.

[12] José Holguín-Veras and Gopal R Patil. A multicommodity integrated freight origin–destination synthesis model. *Networks and Spatial Economics*, 8(2):309–326, 2008.

[13] Chang-i Hua and Frank Porell. A critical review of the development of the gravity model. *International Regional Science Review*, 4(2):97–126, 1979.

[14] Jeffrey A Joines and Christopher R Houck. On the use of non-stationary penalty functions to solve nonlinear constrained optimization problems with ga's. In *Proceedings of the first IEEE conference on evolutionary computation. IEEE world congress on computational intelligence*, pages 579–584. IEEE, 1994.

[15] Woo-Sung Jung, Fengzhong Wang, and H Eugene Stanley. Gravity model in the korean highway. *EPL (Europhysics Letters)*, 81(4):48005, 2008.

[16] Lokesh Kalahasthi, José Holguín-Veras, and Wilfredo F Yushimito. A freight origin-destination synthesis model with mode choice. *Transportation Research Part E: Logistics and Transportation Review*, 157:102595, 2022.

[17] Michael G Kay and Ajithkumar N Parlikad. Material flow analysis of public logistics networks. In *Material Handling Res. Colloq.* Citeseer, 2002.

[18] Michael G. Kay and Donald P. Warsing. Estimating ltl rates using publicly available empirical data. *International Journal of Logistics Research and Applications*, 12(3):165–193, 2009.

[19] Maxime Lenormand, Aleix Bassolas, and José J Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.

[20] Sitian Liu and Yichen Su. The impact of the covid-19 pandemic on the demand for density: Evidence from the us housing market. *Economics Letters*, 207:110010, 2021.

[21] A Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):022812, 2013.

[22] Arakaparampil M Mathai. *An introduction to geometrical probability: distributional aspects with applications*, volume 1. CRC Press, 1999.

[23] KAY Michael. Matlog: Logistics engineering using matlab. *Mühendislik Bilimleri ve Tasarım Dergisi*, 4(1):15–20.

[24] Markus Schläpfer, Lei Dong, Kevin O'Keeffe, Paolo Santi, Michael Szell, Hadrien Salat, Samuel Anklesaria, Mohammad Vazifeh, Carlo Ratti, and Geoffrey B West. The universal visitation law of human mobility. *Nature*, 593(7860):522–527, 2021.

[25] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[26] Rosalyn A Wilson. Transportation in america: Statistical analysis of transportation in the united states. historical compendium 1939-1995. 2002.

[27] Yingxiang Yang, Carlos Herrera, Nathan Eagle, and Marta C González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports*, 4(1):1–9, 2014.

[28] George Kingsley Zipf. The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review*, 11(6):677–686, 1946.