Modeling Crowd-Sourced Spatio-Temporal Flexibility Insights in Origin-Destination Matrices Estimation

Marisdea Castiglione*[1], Guido Cantelmo[2], Ernesto Cipriani[1], and Marialisa Nigro[1]

[1]Department of Civil, Computer Science and Aeronautical Technologies Engineering, Roma Tre University, Via Vito Volterra 62, Rome 00146, Italy
[2]Technical University of Denmark, Bygningstorvet, 116, Kongens Lyngby 2800, Denmark

Short summary

In response to the rapidly evolving urban landscape, there is a growing demand to enhance traditional Origin-Destination Matrices Estimation (ODME) models with new data sources that offer broader perspectives. Crowd-sourced data, in particular, can provide promising avenues for collecting high-resolution data on destination activities, reflecting real mobility patterns. Previous research investigates trip motivations and the varying travel flexibility associated with different activities, leveraging real-world crowd-sourced data like Floating Car Data (FCD) and Google Popular Times (GPT).

To address the need to integrate these insights into ODME models, this paper introduces the Flex-GLS approach, an extension of the GLS model that accounts for multiple demand components characterized by spatio-temporal flexibility metrics derived from crowd-sourced data. It aims to offer a more precise representation of travel demand by integrating both temporal and spatial flexibility dimensions. Benchmarking against the traditional GLS model highlights the potential of Flex-GLS in enhancing ODME accuracy, providing valuable insights into urban travel dynamics.

Keywords: Crowd-sourced data; Floating Car Data; OD Matrices Estimation; Spatio-temporal Flexibility; Trip Purpose.

1 Introduction

In response to the challenges posed by the rapidly evolving urban environment, there is an emerging need to supplement traditional Origin-Destination Matrices Estimation (ODME) models with data sources that can offer broader and more insightful perspectives (Cantelmo et al., 2014). While traditional fixed-location data collection tools have been foundational in establishing reliable metrics of traffic flows, such as vehicle counts, speeds, and densities, they often do not fully encompass the complexity of urban mobility, leading to an incomplete representation of the multifaceted nature of travel demand (Carrese et al., 2017).

Emerging technologies and in particular crowd-sourced data – which includes mobile phone data, GPS-based data, and social media analytics – offer promising avenues for gathering high-resolution data that reflects the actual travel patterns of urban travelers and can significantly enrich ODME models. In particular, the integration of location-based crowd-sourced data (Timokhin et al., 2020) into ODME models can offer significant insights into the activities users engage in at destination or the purpose of their trips.

In an earlier work Castiglione et al. (2024), the authors investigated the motivations behind people's trips, estimating how travel flexibility varies in relation the nature of activities at their destinations, leveraging crowd-sourced data such as Floating Car Data (FCD) and Google Popular Times (GPT). The evaluation of flexibility parameters computed for each user within the FCD dataset has revealed variability both across and within activity types, as well as over different time frames. This variability enhances the depth of the analysis, allowing for detailed estimations of spatio-temporal flexibility for different components of travel demand.

Four demand components C have been identified, each associated with a specific level of temporal and spatial flexibility, represented by a set of n_t sample OD matrices. These matrices are derived through the aggregation of FCD trips, categorized based on comparable values of spatio-temporal flexibility. The evidence from earlier research highlights the necessity of integrating the insights obtained from crowd-sourced data into the framework of traditional dynamic ODME models, such as the Generalised Least Squares (GLS) (Cascetta et al., 1993) model.

This paper introduces the Flex-GLS approach, a novel extension of the GLS model that is designed to account for multiple demand components characterised by spatio-temporal flexibility metrics derived from real-world, crowd-sourced data. This model aims to offer a more accurate representation of travel demand by integrating both temporal and spatial flexibility dimensions, thereby better reflecting the complex dynamics of urban travel.

The paper is organized as follows: the 'Methodology' section describes the Flex-GLS model, detailing its theoretical foundations based on the concepts of Temporal and Spatial Flexibility. The 'Results and Discussion' section presents numerical applications of the model, analyzing its performance against the standard GLS. Finally, the 'Conclusions' section summarizes the findings and suggests future research directions.

2 Methodology

Given $n_t \times n_C$ sample Origin-Destination (OD) matrices, each cell of these matrices represent trips from origin O to destination D within a time interval t belonging to a component of travel demand C, obtained from the aggregation of crowd-sourced data based on shared spatio-temporal flexibility values σ_C, as done in Castiglione et al. (2024).

Here, Temporal Flexibility (TF) indicates the degree to which an individual can adjust the timing of their activities, whereas Spatial Flexibility (SF) denotes the extent to which an individual can vary the locations of their activities. In this context, the traditional Generalized Least Squares (GLS) formulation, as referenced by (Cascetta et al., 1993), is extended to encompass multiple demand components C, each characterized by its unique spatio-temporal flexibility distribution σ_C. The modified objective function formulation is provided as follows:

$$d^* = \arg\min_{d^*} \left(\sum_t \left(\sum_l w_l \cdot (v_l(d^*) - \hat{v}_l)^2 + \sum_{od} \sum_C w_C \cdot (d^*_{od,C} - \hat{d}_{od,C})^2 \right) \right) \tag{1}$$

Here, $v_l(d^*)$ denotes simulated traffic flows from estimated demand d^*, against observed traffic counts $\hat{v}_l$. The demand matrix d^*, segmented into C components where $d^*_{od,C}$ indicates demand for each origin-destination pair per component, while $\hat{d}_{od,C}$ is the seed matrix for each demand component obtained from the classified FCD in Castiglione et al. (2024). The weights w_l and w_C are assigned based on the inverse of traffic counts and demand component variances, respectively.

Extending the GLS method to encompass multiple demand components, while conceptually simple, adds significant complexity to the estimation process, particularly in large urban networks. This approach greatly increases the number of variables in the already intricate ODME problem, escalating computational demands.

To manage this complexity, the Flex-GLS takes advantage of the fact that the total demand d^* is the sum of all the individual components $d^*_{od,C}$, leveraging conditional probabilities to model them as a unified variable. This method keeps variable count manageable and ensures computational feasibility. In the Flex-GLS model, the demand for each origin-destination (od) pair and time interval (t) is estimated, then, each individual demand component is adjusted using temporal and spatial flexibility metrics.

Like the traditional GLS, the Flex-GLS model is an iterative optimization procedure that begins with an Initialization phase to set parameters for the ODME model, followed by the Assignment phase for allocating the estimated demand across the network. Given the observed traffic counts, the model applies a Gradient Descent algorithm, guided by the function in Equation (1), to fine-tune the demand.

Demand components are defined as:

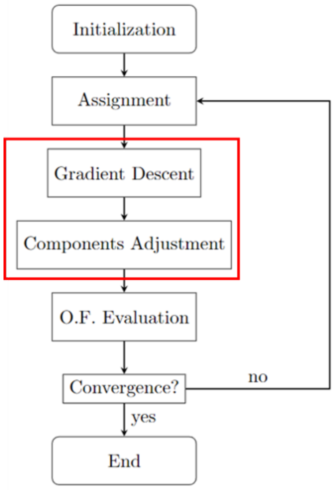

Figure 1: Flex-GLS Flowchart

$$d^*_{C,t,od} = p^*_C(t) \times d^*_{t,od} \tag{2}$$

$$p_C(t) = \frac{d_{C,t}}{\sum_t d_{C,t}} \tag{3}$$

Where $p^*_C(t)$ is derived from the classified FCD sample OD matrices. The objective function is reformulated as:

$$d^* = \arg\min_{d^*} \left(\sum_t \left(\sum_l w_l \cdot (v_{l,t}(d^*) - \hat{v}_{l,t})^2 + \sum_{od} \sum_C w_C \cdot (p^*_C(t) \times d^*_{t,od} - \hat{d}_{C,t,od})^2 \right) \right) \tag{4}$$

The Flex-GLS uses a gradient descent algorithm to minimize the objective function. At each iteration k, the demand is updated as follows:

$$d^{*(k+1)}_{t,od} = d^{*(k)}_{t,od} - \alpha \cdot \frac{\partial}{\partial d^*_{t,od}} \left(\sum_{\text{od}} \sum_t w_{l,t} \cdot \left(v_{l,t}\left(d^*\right) - \hat{v}_{l,t}\right)^2 + w_C \cdot \left(p^*_C(t) \cdot d^*_{t,od} - \hat{d}_{C,t,od}\right)^2 \right) \tag{5}$$

with α as the learning rate.

The first term of the gradient, which accounts for discrepancies in traffic flows, adheres to the traditional GLS approach. As for the second component of the gradient, it is important to recognize that, as the demand components d^*_C are not independent variables, but rather a function of the total demand d^* and the corresponding proportion for component C at time t, denoted as $P^*_C(t)$, the gradient descent update rule at iteration k for each demand component is formulated as:

$$\frac{\partial}{\partial d^*} \left(w_C \cdot \left(P^*_C(t) \cdot d^*_{t,od} - \hat{d}_{C,t,od} \right)^2 \right) = 2 \cdot \frac{1}{\sigma_C} \cdot \left(P^*_C(t) \cdot d^*_{t,od} - \hat{d}_{C,t,od} \right) \cdot P^*_C(t) \tag{6}$$

The weights w_C are modulated to be the inverse of the variance of each component's demand ($w_C = \frac{1}{\sigma_C}$), thereby emphasizing the impact of variations in the more rigid demand components within the objective function. This is due to their smaller variance, which inherently discourages large fluctuations in the less flexible demand components.

Post-estimation, the model enters the demand Components Adjustment phase to recalibrate $p^*_C(t)$ for each demand component based on flexibility metrics, to ensure alignment with the newly estimated demand. If the evaluation of the objective function achieves satisfactory convergence, the process concludes; otherwise, it cycles back to the assignment phase. The overall process is illustrated in Figure 1, which highlights the Gradient Descent and Components Adjustment steps as they represesent the phases that diverge from the traditional GLS approach.

After each Gradient Descent step, the Flex-GLS refines the individual demand components through a constrained Maximum Likelihood Estimation (MLE) problem, utilizing prior probabilities $P_C(t)$

and variances σ_C^2 based on seed FCD data. Updated at each iteration, these probabilities align with the total demand d^*. The model's total likelihood, L_{total}, is computed as the logarithm of individual component likelihoods.

The MLE constraints in Flex-GLS aim to maintain consistency with the observed data. The first constraint (Eq. 7) ensures the sum of probabilities $P_C^*(t)$ for all components C equals one at each od pair and time interval t:

$$\sum_C P_C^*(t) = 1, \quad \forall t \in T, \forall od \tag{7}$$

A second constraint (Eq. 8) maintains the temporal consistency of demand components' proportions, both pre- and post-estimation:

$$\frac{\sum_{t\in T} d_{C,t,od}}{\sum_{t\in T} d_{t,od}} \approx \frac{\sum_{t\in T} d^*_{C,t,od}}{\sum_{t\in T} d^*_{t,od}} \quad \forall C, \forall od \tag{8}$$

The model accommodates variation ϵ in demand component ratios to reflect flexibility, with lower ϵ for rigid components and higher for flexible ones. This Temporal constraint dictates that the demand components proportions within time interval t can be adjusted according to the flexibility levels, but maintain overall consistency within the broader temporal window T.

For instance, we can consider the example of commuters whose usual departure time is 8:00 am. Their temporal flexibility might allow to change their departure time within a narrow window, such as 7:45 to 8:15, as opposed to users traveling for different purposes (e.g. shopping) who may exhibit a broader flexibility window. This highlights the importance of an accurate definition of the time window T in capturing variations in travel behavior.

Similarly, Spatial Flexibility involves re-distributing the demand generated from one origin O within the same time interval across various destination, maintaining consistency in generated demand component proportions. The constraint for Spatial Flexibility is as follows:

$$\frac{\sum_d d_{C,t,od}}{\sum_d d_{t,od}} \approx \frac{\sum_d d^*_{C,t,od}}{\sum_d d^*_{t,od}} \quad \forall C, \forall t \tag{9}$$

In the Flex-GLS model, Temporal and Spatial Flexibility are treated as complementary. The constraints for the joint spatio-temporal MLE problem are redefined as follows:

$$\begin{cases} \sum_C P_C^*(t) = 1, \quad \forall t \in T, \forall od \\ \frac{\sum_{t\in T}\sum_d d^*_{C,t,od}}{\sum_{t\in T}\sum_d d_{t,od}} \approx \frac{\sum_{t\in T}\sum_d d^*_{C,t,od}}{\sum_{t\in T}\sum_d d^*_{t,od}}, \quad \forall C, \forall t \in T, \forall d \end{cases} \tag{10}$$

The demand component constraints in the Flex-GLS model can be visualized as a matrix segmented into blocks across od and t dimensions. The constraints for temporal and spatial flexibility determine the size and overlap of these blocks: rigid demand components are represented by smaller, closely overlapping blocks, exemplified by the limited departure time or destination choice flexibility . In contrast, blocks representing flexible demand components are larger, spanning multiple time intervals, indicative of looser constraints and a broader range of departure time flexibility as well as destination choices.

3 Results and discussion

A systematic comparison is conducted for benchmarking purposes, to assess the Flex-GLS model against the conventional GLS model across various scenarios. The underlying hypothesis posits that Flex-GLS, as a generalization of GLS, offers improved accuracy in estimation, particularly when the demand components' proportions obtained through crowd-sourced data aligns closely with reality.
The benchmarking tests use a network with one origin and two destinations (Figure 3), focusing on spatial flexibility across different demand components in a single time interval. This setup allows for the exploration of three distinct comparative scenarios:

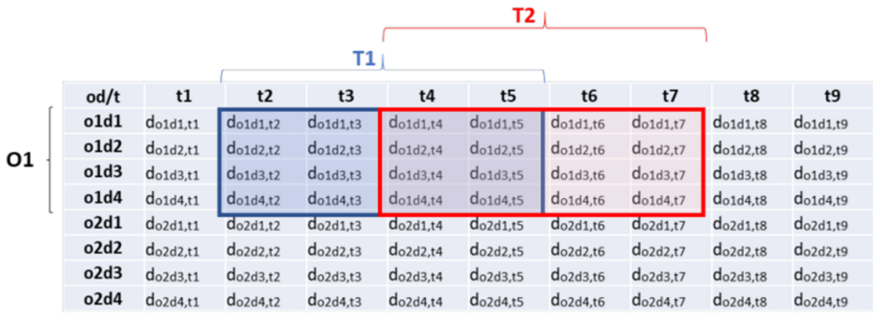

Figure 2: Representation of the constraints blocks within a demand component matrix

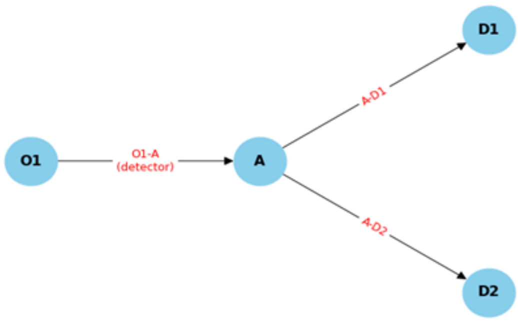

Figure 3: Benchmarking Analysis Network

1. Scenario 1: High congruence between seed and real demand matrices as well as in component ratios.
2. Scenario 2: Divergence in total seed demand from actual demand, but similarity in component proportions.
3. Scenario 3: Close match in total seed and real demands, but significant differences in component ratios.

Each scenario offers insights into the models' performance and adaptability, using a "real" demand as a baseline for comparison, which, while not directly observable, represents the actual demand generating the traffic counts on the network.

Table 1: Real Demand and Detected Counts

OD Pair	Real Demand	Rigid Real Demand	Flex Real Demand	Detected Counts
OD1-1	50	45	5	150
OD1-2	100	45	55	

The seed demand matrix, as outlined in Table 1, serves as the foundation for each scenario, constructed by introducing perturbations to the real demand (or its demand components) with differing levels of noise.

Scenario 1: High Reliability of both Total and Component Seed Demand

In this scenario, seed demand and its components closely match the actual demands with minor perturbations introduced: up to ±3% for the rigid component and ±10% for the flexible component to reflect their inherent stability and variability, respectively.

Table 2: Scenario 1: Seed Demand and Initial Flows

OD ID	Seed Demand	Rigid Seed Demand	Flexible Seed Demand	Initial Flows
OD1-1	49	44	5	140
OD1-2	91	43	48	

In this scenario, the Flex-GLS model demonstrates enhanced performance compared to GLS. This is evidenced not just in terms of accurately reproducing network flows but also in effectively estimating the total demand. Additionally, it proves more adept in estimating the individual rigid and flexible demand components, as indicated by the comparative RMSE values summarized in the following table:

Table 3: Scenario 1: GLS and Flex-GLS RMSE Comparison

RMSE	GLS	Flex-GLS
Detected Counts vs. Simulated Flows (Initial)	9.200838	9.200838
Detected Counts vs. Simulated Flows (Final)	3.067084	1.086609
Real vs. Seed Demand	5.983042	5.983042
Real vs. Estimated Demand	4.121249	3.316182
Seed vs. Estimated Demand	3.066877	4.094755
Rigid Real vs. Rigid Seed	1.584283	1.584283
Rigid Real vs. Rigid Estimated	1.603394	0.847237
Flexible Real vs. Flexible Seed	4.435897	4.435897
Flexible Real vs. Flexible Estimated	3.282010	2.993162

The comparative RMSE values further illustrate the Flex-GLS model's superior ability to adjust demand estimations accurately, showcasing its nuanced handling of demand component variability and stability.

Scenario 2: Low Reliability of Total Seed Demand and High Reliability of Component Demand

Scenario 2 tests the Flex-GLS model's accuracy in component-wise demand estimation under conditions where total seed demand significantly deviates from actual demand, yet component structures remain aligned with reality. Real demand components are adjusted with random noise ($\pm 3\%$ for rigid, $\pm 10\%$ for flexible components) before inflating total seed demand by 30%, maintaining component ratio integrity.

Table 4: Scenario 2: Seed Demand and Initial Flows

OD ID	Seed Demand	Rigid Seed Demand	Flexible Seed Demand	Initial Flows
OD1-1	70	57	13	208
OD1-2	138	61	77	

The Flex-GLS model demonstrates improved performance over the GLS in flow reproduction and demand estimation, including individual demand components.

Table 5: Scenario 2: GLS and Flex-GLS RMSE Comparison

RMSE	GLS	Flex-GLS
Detected Counts vs Simulated Flows (Initial)	58.280870	58.280870
Detected Counts vs Simulated Flows (Final)	19.427832	5.855703
Real Demand vs Seed Demand	29.968099	29.968099
Real Demand vs Estimated Demand	11.970053	4.311141
Seed Demand vs Estimated Demand	19.426519	26.490912
Rigid Real Demand vs Rigid Seed Demand	13.842191	13.842191
Rigid Real Demand vs Rigid Estimated Demand	4.009714	2.703550
Flexible Real Demand vs Flexible Seed Demand	16.800260	16.800260
Flexible Real Demand vs Flexible Estimated Demand	8.693669	6.433310

This showcases the Flex-GLS model's robustness and precision in estimating demands, even when facing significant discrepancies between seed and actual demands, emphasizing its capability to effectively reconcile component-wise estimations.

Scenario 3: High Reliability of Total Seed Demand, Low Reliability of Component Demand

Scenario 3 evaluates the model's performance with accurate total seed demand but inaccurate component distributions, testing the Flex-GLS model's precision in estimating demand components under these conditions.

Table 6: Scenario 3: Seed Demand and Initial Flows

OD ID	Seed Demand	Rigid Seed Demand	Flexible Seed Demand	Initial Flows
OD1-1	49	8	41	151
OD1-2	102	7	95	

The RMSE values for the GLS and Flex-GLS in Scenario 3 are as follows:

Table 7: Scenario 3: GLS and Flex-GLS RMSE Comparison

RMSE	GLS	Flex-GLS
Detected Counts vs Simulated Flows (Initial)	2.233991	2.233991
Detected Counts vs Simulated Flows (Final)	0.744697	0.250565
Real Demand vs Seed Demand	1.754590	1.754590
Real Demand vs Estimated Demand	1.403406	12.125056
Seed Demand vs Estimated Demand	0.744647	10.842700
Rigid Real Demand vs Rigid Seed Demand	37.297531	37.297531
Rigid Real Demand vs Rigid Estimated Demand	37.385913	36.649881
Flexible Real Demand vs Flexible Seed Demand	38.471518	38.471518
Flexible Real Demand vs Flexible Estimated Demand	37.815052	36.087185

In this scenario, the Flex-GLS model underperforms, leading to high RMSE values in demand comparisons, indicating substantial inaccuracies in the estimation. This scenario, however, wants to represent an extreme case, emphasizing that the Flex-GLS is most effective when reliable data on the structure of individual demand components, such as those derived from crowd-sourced data, is available. If such data is unreliable or absent, a hybrid approach incorporating elements of both GLS and Flex-GLS might be advisable, particularly in less extreme scenarios.

4 Conclusions

This paper introduces a new extension of the GLS ODME framework to include crowd-sourced data. The Flex-GLS model is tested on three scenario-based evaluations, which reveal the applicability of the model under different conditions.

When the seed demand matrices closely mirror the real demand, as in Scenario 1, the Flex-GLS model demonstrates its capability to harness the structural insights of the data, yielding estimations with remarkable accuracy to the observed traffic conditions. The model's superior performance in this scenario underscores its potential in contexts where the empirical data is robust and the ratios between demand components are reliable.

Scenario 2 introduces a challenge where the total seed demand is very perturbed, yet the Flex-GLS model adeptly adjusts the demand estimations, attesting to its resilience when the individual components structure is mantained. This scenario exemplifies the model's robustness against large-scale deviations in overall demand while preserving the integrity of component-specific estimations.

However, the model's sensitivity to the reliability of component structures becomes apparent in Scenario 3. When the component demands are significantly misaligned with actual demands, the Flex-GLS model's estimations become less reliable, resulting in elevated RMSE values despite the seed demand being very close to the real demand. This outcome cautions against the Flex-GLS's use in the absence of dependable data on the structure of the individual demand components. However, it is important to note that the insights obtainable from crowd-sourced data align with these requirements.

Ultimately, the outcomes from these scenarios guide the strategic selection of the most appropriate modeling approach, advocating for the Flex-GLS when detailed, reliable crowd-sourced data is available, and for the standard GLS in less certain data environments.

References

Cantelmo, G., Cipriani, E., Gemma, A., & Nigro, M. (2014, June). An Adaptive Bi-Level Gradient Procedure for the Estimation of Dynamic Traffic Demand. *IEEE Transactions on Intelligent Transportation Systems*, *15*(3), 1348–1361. Retrieved 2024-01-17, from `https://ieeexplore.ieee.org/abstract/document/6740860` (Conference Name: IEEE Transactions on Intelligent Transportation Systems) doi: 10.1109/TITS.2014.2299734

Carrese, S., Cipriani, E., Mannini, L., & Nigro, M. (2017, August). Dynamic demand estimation and prediction for traffic urban networks adopting new data sources. *Transportation Research Part C: Emerging Technologies*, *81*, 83–98. Retrieved 2024-01-17, from `https://linkinghub.elsevier.com/retrieve/pii/S0968090X17301432` doi: 10.1016/j.trc.2017.05.013

Cascetta, E., Inaudi, D., & Marquis, G. (1993, November). Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts. *Transportation Science*, *27*(4), 363–373. Retrieved 2024-01-17, from `https://pubsonline.informs.org/doi/10.1287/trsc.27.4.363` doi: 10.1287/trsc.27.4.363

Castiglione, M., Cantelmo, G., Cipriani, E., & Nigro, M. (2024). From trip purpose to space-time flexibility: A study using floating car data and google popular times. *TRANSPORTMETRICA B: Transport Dynamics*. (Under review)

Timokhin, S., Sadrani, M., & Antoniou, C. (2020, August). Predicting Venue Popularity Using Crowd-Sourced and Passive Sensor Data. *Smart Cities*, *3*, 818–841. doi: 10.3390/smartcities3030042