Modeling Checkpoint-Based Movement with the Earth Mover's Distance*

Matt Duckham[1], Marc van Kreveld[2], Ross Purves[3],
Bettina Speckmann[4], Yaguang Tao[1], Kevin Verbeek[4], and Jo Wood[5]

[1] School of Science, RMIT University, Melbourne, Australia.
matt.duckham@rmit.edu.au s3553285@student.rmit.edu.au
[2] Dept. of Computing and Information Sciences, Utrecht University, the Netherlands.
m.j.vankreveld@uu.nl
[3] Dept. of Geography, University of Zurich, Switzerland. ross.purves@geo.uzh.ch
[4] Dept. of Mathematics and Computer Science, TU Eindhoven, the Netherlands.
{k.a.b.verbeek b.speckmann}@tue.nl
[5] Dept. of Computer Science, City University London, UK. J.D.Wood@city.ac.uk

Abstract. Movement data comes in various forms, including trajectory data and checkpoint data. While trajectories give detailed information about the movement of individual entities, checkpoint data in its simplest form does not give identities, just counts at checkpoints. However, checkpoint data is of increasing interest since it is readily available due to privacy reasons and as a by-product of other data collection. In this paper we propose to use the Earth Mover's Distance as a versatile tool to reconstruct individual movements or flow based on checkpoint counts at different times. We analyze the modeling possibilities and provide experiments that validate model predictions, based on coarse-grained aggregations of data about actual movements of couriers in London, UK. While we cannot expect to reconstruct precise individual movements from highly granular checkpoint data, the evaluation does show that the approach can generate meaningful estimates of object movements.

1 Introduction

Throughout the years, interest in spatial data has shifted from static planar maps, to space-time [19] and 3D GIS [1], and to movement data [21],[29]. The study of movement data has grown explosively due to the availability of tracking devices and their increased quality. Movement is essential for modeling many types of spatial interaction, one of the central concepts in spatial analysis.

Movement data is often available in the form of *trajectories:* sequences of time-stamped locations acquired through GPS or other devices that can determine the location of an individual entity. There are a host of computational

* B. Speckmann and K. Verbeek are supported by the Netherlands Organisation for Scientific Research (NWO) under project nos. 639.023.208 and 639.021.541, respectively. This paper arose from work initiated at Dagstuhl seminar 12512 "Representation, analysis and visualization of moving objects", December 2012. The authors gratefully acknowledge Schloss Dagstuhl for their support.

tools to analyze trajectories, for example, to determine similarity, to cluster, or to find specific patterns in the trajectories (such as flocks or leadership) or the underlying space (like hotspots) [9],[16].

Recently, a different type of movement data has become of increasing interest, namely *checkpoint data* [3,4],[6],[11],[25,26],[28]. Here the entities themselves need not be equipped with GPS, but rather their presence at a location or neighborhood is recorded by a stationary sensor. Such sensors include street cameras counting passing pedestrians, check-in gates at metro stations, inductive loops counting cars, RFID sensors in mass participation sporting events, and mobile phone cell towers and wifi access points counting the number of connections in their vicinity. The resulting type of movement data is typically either anonymous or anonymized before being made available for analysis. Hence, frequently the only data available is counts of entities at certain times or in certain intervals.

Checkpoint data is usually much less information-rich than trajectory data. This is partly due to the typically coarse spatial granularity of fixed checkpoint locations, but also due to the lack of heading, speed, chosen route, and stops that are not recorded nor so easily derived from aggregate counts. We can identify several types of checkpoint data based on the spatial extent of acquisition of the data (point-based or area-based) and the movement space (network or more general). Examples of the resulting four classes are given in Table 1.

Table 1. Examples of various types of checkpoint data.

	Network movement	Areal movement
Point-based check (cameras, gates, inductive loops)	Road traffic, subway	Indoor movement (airport, hall)
Area-based check (cell towers, satellite)	Pedestrians (street)	Pedestrians (square, park)

The coarse-grained aspect of the data makes it suitable only for coarse-grained pattern analysis. Perhaps the most important one of these patterns is global *flow* of entities. But since no identity, heading, or speed data is available, flow must be reconstructed from the counts. Reconstruction of flow can be based on any of various *spatial interaction models*. Spatial interaction models describe the flow of people, goods, infections, or information between locations in geographic space, and are therefore studied in various fields of geography.

In this paper we assume a tessellated geographic space and a number of time stamps as a model for area-based checkpoint data. At each time stamp or snapshot, we have a count of the number of entities in each region of the tessellation (termed "temporal checkpoints" in [25], akin to a function from time to a spatial field). Such data may arise from mobile phone connection counts in cell tower regions, for example, aggregated over time intervals.

We will study the possibilities of reconstructing flow consistent with this data using the *Earth Mover's Distance* [23], a well-known measure for capturing the distance (or its inverse, similarity) between two images or weighted point sets. It has also been used in GIS for similarity assessment (see, for example,

[10],[15],[17]). Let $R = \{(r_1, w_1), \ldots, (r_n, w_n)\}$ be a set of n tuples consisting of points r_i and corresponding weights w_i. Let $W = \sum w_i$. Similarly, let $B = \{(b_1, v_1), \ldots, (b_m, v_m)\}$ be a set of m tuples, and let $V = \sum v_i$. The Earth Mover's Distance between R and B is defined if $W = V$, and is the minimum total effort to transport all the weight from R to B. The effort to transport weight w from a point r to a point b is defined as $w \cdot \text{dist}(r, b)$, where $\text{dist}(r, b)$ is a distance measure, for example the Euclidean distance. The Earth Mover's Distance is a metric, also known as the Wasserstein metric. Since the total weight in R and B is the same, we must transport all weight from R to give all points of B the correct weight. Any point in R can give its weight to multiple points in B, and any point in B may receive its weight from one or more points in R. Therefore, a minimum effort transportation corresponds to a flow from R to B.

Reconstructing flows allows us to make effective visualizations including OD maps [27] and flow maps [2],[5]. Figure 1 shows a typical output of our model estimating flows of people based on granular mobile phone data.

Results and Organization. In Section 2 we overview spatial interaction models and argue that the Earth Mover's Distance is suitable for reconstructing flow from checkpoint data. We recap a linear-programming formulation to compute the Earth Mover's Distance. In Section 3 we use the Earth Mover's Distance to reconstruct flow in typical scenarios like mobility in a city. We show that environmental situations like obstacles (rivers) and metro stations can easily be incorporated by adapting the objective function and constraints of the linear program. In Section 4 we analyze the success of the Earth Mover's Distance to reconstruct flows. To this end we evaluate our approach using data about real trajectories of couriers in London, UK, by converting them to tessellated counts at time stamps and then trying to reconstruct the flows present in the origi-

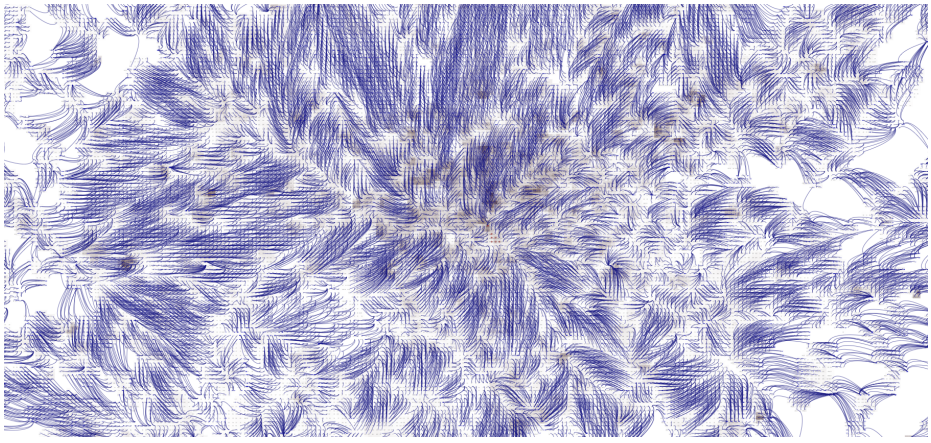


Fig. 1. Estimated flows of people between 9:05 and 9:10am in central London, 3rd June 2012. Flow estimations based on least cost movement between mobile telephone density surfaces over the 5 minute period.

nal trajectories. Section 5 summarizes the contribution of this work, as well as indicating further possibilities and improvements for future work.

2 Spatial Interaction and the Earth Mover's Distance

Spatial interaction models of flow are commonly associated with the gravity model [8],[20],[22], which in its original form relates the trade flow F_{ij} between two countries i, j using their economic masses M_i and M_j and their distance d_{ij}:

$$F_{ij} = c \cdot \frac{M_i^{\beta_1} M_j^{\beta_2}}{d_{ij}^{\beta_3}} \ ,$$

where $c, \beta_1, \beta_2, \beta_3$ are constants. The distance may be influenced by the cost of transportation but also by trade barriers. Many extensions of the gravity model have been described, taking into account more factors or compensating for weaknesses. Besides economics, the gravity model is also popular in transportation, migration, and mobility modeling. Other spatial interaction models include the radiation model [24] and Huff's probabilistic model [13,14].

While these models could be used to model movement in checkpoint data, the Earth Mover's Distance [23] (EMD) has potential advantages. The other models aim to represent global patterns of interaction, established over long time periods (over which small variations are smoothed out), and focus on economic principles such as supply and demand. There is little reason to believe that such models would work well for reconstruction of movement based on checkpoint data, which has a much finer time resolution and may vary rapidly in both time and space. Furthermore, models like the gravity model attempt to explain the degree of interaction based on (economic) masses without taking local patterns into account. A gravity function can be fitted to the data, but such a function will be global and apply to the whole grid. Geographically weighted regression approach [18] has been taken recently to support local spatial interaction modeling. While more location specific parameters were introduced to reach a better fit, existing flow data is required in training the models. Our objective is to reconstruct deviations from global movement behavior, or random patterns, and detect local trends of movement that exist in specific areas at specific times. For this we use one of the simplest possible models, the EMD. Importantly the EMD conserves mass in flow, although we purposefully adapt it to account for loss or gain of mass, for example, because of sensor error or movements not detected by sensors.

We consider a specific instance of checkpoint data where we have counts at time steps $t_1, \ldots, t_s$ at all checkpoints. For descriptive purposes we assume that the checkpoints provide counts in regions of a regular grid. When a grid of counts at time t_i and a grid of counts at time t_{i+1} are known, we can infer movement from entities in cells at time t_i to cells at time t_{i+1}, see Fig. 2. In particular, if some cell c contains 10 entities at time t_i and 6 entities at time t_{i+1}, we are certain that at least 4 entities have left the cell. Possibly, all 10 entities have left

and 6 other entities appeared. It is also possible that yet other entities passed all the way through cell c between times t_i and t_{i+1} and were never counted.

Checkpoint data does not allow us to completely reconstruct flow, since, for example, it is difficult to identify flow between two cells of the same magnitude, because they cancel out. However, we can still hope to determine flows at a somewhat more global level if there is a trend. To this end, we make an assumption of minimum cost movement. We do not claim that this is realistic, but it does provide a lower bound on the total flow. Minimum cost flow can be derived from the EMD, as described in the introduction. We let the location of an entity be the center of the cell the entity is in. So an entity sits in the same cell at time t_i and t_{i+1} has exactly the same location despite that it might moved slightly. When a minimum cost flow lets entities move to the same cell as the one they started, the cost of the movement is zero because the movement distance is zero. Movement of entities to an adjacent cell has cost equal to the product of the cell size and the number of entities moving. In Fig. 2 there are two minimum flow solutions.

The minimum cost flow problem can be formulated as a linear program. Here the flow from a cell j at t_i to a cell k at time t_{i+1} becomes a variable F_{jk}. The objective function, to be minimized, is the summation of all flows times the distance:

$$\sum_{j,k} F_{jk} \cdot d_{jk} \tag{1}$$

where the distance is assumed to be the distance between the cell centers. To ensure that the flow transports the correct numbers, we use constraints. They come in three types:

- Non-negativity constraint: $F_{jk} \geq 0$ for all j, k
- Origin constraint: $\sum_k F_{jk} =$ count of cell j at time t_i, for all j
- Destination constraint: $\sum_j F_{jk} =$ count of cell k at time t_{i+1}, for all k

In principle, no flow is negative, the whole count must exit each cell at time t_i, and the resulting count at each cell has arrived at time t_{i+1}. We can replace the two equalities by inequalities and obtain a linear program with the same solution.

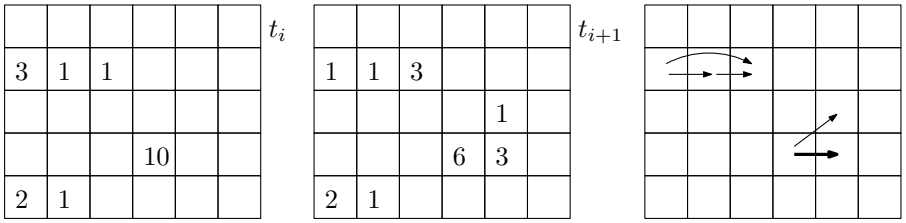


Fig. 2. Grid with counts at times t_i and t_{i+1}, and a possible flow indicated to the right by arrows. In another minimum cost movement, two of the three entities at the top left moved two cells to the right, and the one entity in the middle did not move.

3 Modeling and Computation of Flow in Specific Cases

To demonstrate the versatility of linear programming to compute the EMD flow, we show how to incorporate various situations in a natural way. We consider flow in an urban environment based on mobile phone data and a time interval of 5 minutes. This is a typical situation in practice. It allows identification of the main flows during morning and evening rush hours, flow during big events, and generally flow patterns at different times. We can imagine a grid of, for instance, 20×20 cells, each of 100×100 meters.

Urban movement from area-based counts. The basic computation of flow using the EMD follows the three linear programming constraints given in the previous section. The EMD is in principle mass preserving, but we can expect that in our situation of urban movement there will be different total counts at times t_i and t_{i+1}. There are two main reasons for this:

- People at the edges of the area of interest move to the outside, or people just outside the area of interest move inside. We can assume that this movement influences the counts in the cells close to the boundary.
- People can at any time switch on or off their device, and they may also lose connection or acquire a connection.

To incorporate the former we extend the grid with an extra ring of cells surrounding the original grid, see Fig. 3. The extra ring does not have data, so there are no counts for these cells. In our model we allow these cells to produce extra entities moving into the core grid, or take up entities departing from the core grid. This models the boundary effects in a simple and elegant manner as a (potentially infinite) sink/source.

To incorporate the latter we allow entities in every cell to disappear or appear in a count. Since we prefer to "explain" changing numbers by movement, the cost of appearing or disappearing will be significantly higher than that of movement. Technically, we add one extra "cell" to t_i and to t_{i+1}, which does not have a location. The extra cell in t_i (and t_{i+1}) allows movement of any number to (from)

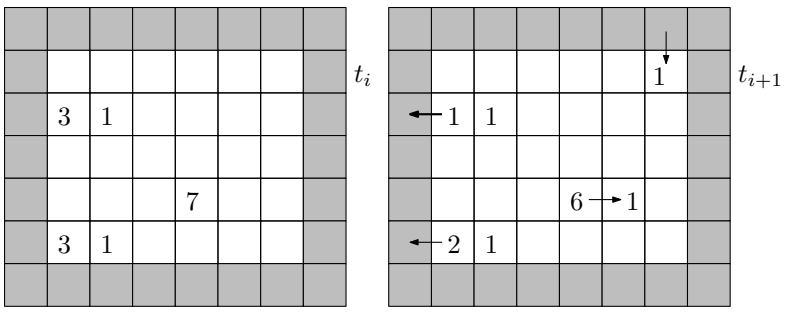


Fig. 3. An extra ring of cells (grey) around the core grid (white) allows us to model movement of entities to and from the outside area.

any cell in t_{i+1} (and t_i) at the same, high cost per unit. That is, we set d_{cj} and $d_{kc'}$ to high values in Eq. (1) when c and c' are the extra cells.

In many big cities, a major reason for losing cell tower connection is going underground to take a subway. This can be incorporated easily in our model. Grid cells that contain an entrance to the subway have a lower cost (captured in d_{jk}) of appearing and disappearing. The same applies to cells from which a subway entrance can be reached, naturally incorporating the distance between the cell and the subway entrance.

Another common feature is the presence of obstacles in a city, like a river or a stretch of train tracks that does not have crossings. Such situations can cause two nearby grid cells to be much further apart by travel distance than by Euclidean distance. So again, we need only change the distance function d_{jk} in Eq. (1) to accommodate for the increased distance. It is reasonable to use the *geodesic distance,* the length of the shortest path that does not cross obstacles, as the altered distance.

In the model we can choose to favor many small movements over fewer larger movements or vice versa. With the linear conversion of distance to cost in Eq. (1) we observe that five unit-distance movements cost as much as one movement over five units. By raising d_{jk} to a power γ we can favor smaller movements by setting $\gamma > 1$ or larger movements by setting $\gamma < 1$. The parameter γ is closely related to β_3 in the gravity model and corresponds to the concept of *distance decay.* Also note that the LP remains linear in its variables, so this adaptation does not influence efficiency.

We observe that it is generally not possible in our scenario to get from any cell to any other cell in a given time interval. By assuming a maximum travel speed in the city, we can limit the number of cells that can be reached from any cell. This has a positive effect on both the resulting flow (we forbid long-distance, unrealistic flows) and on the efficiency. Since the LP has a flow variable F_{jk} for every cell pair j at t_i and k at t_{i+1} between which flow is possible, we can reduce the number of variables drastically this way.

Finally, we observe that the assumption of a grid is not necessary for EMD and its LP-based algorithm. For any partition into regions we can use a representative point inside (the cell tower location) instead of the grid cell center.

Other movement. Movement monitored by gates or cameras leads to point-based counts rather than area-based counts. With toll gates on highways and with check-in gates of subways, we know the direction and precise count of entities accessing a particular area; with cameras this is less precise. Previous research on traffic management in combination with checkpoints concentrated on toll gate placement and pricing [7,11,28], travel time estimation [3,26], or traffic flow modeling in general, see, for example, [12].

We briefly discuss movement described by point-based counts, because it is considerably different from movement described by area-based counts. We assume a network is given with certain positions where check-in and check-out is possible. Again our objective is to determine flow, which is closely related to matching up in-flow of the network with out-flow. For example, if there is a large

check-in count at checkpoints a and i in Fig. 4, and a large check-out count later at checkpoints f and d, it is interesting to try and determine if entities mostly went from a to d and from i to f, or if they mostly went from a to f and from i to d.

With area-based checkpoints every entity – in theory – is counted once by a checkpoint at any time. With point-based checkpoints, time plays a different role. To be able to compute a matching also for point-based checkpoint data, we can generate check-in counts and check-out counts in 5-minute intervals. This results in two sets of weighted points, where the points are a combination (c, i) of checkpoint c and time interval t_i, and the weights are the corresponding counts. Thus, we can again use the EMD to reconstruct flow. This results in flows of the form $F_{c,i,c',j}$, where $F_{c,i,c',j}$ describes the potential flow from any check-in point c at time interval t_i to any check-out point c' at time interval t_j. In our LP, we need a variable for $F_{c,i,c',j}$ only if $j \geq i$, or more generally, if the trip from c to c' is possible in $t_j - t_i$ time, plus the sampling interval. It is natural to use the typical travel times between checkpoints to obtain the most likely matching. For a particular flow $F_{c,i,c',j}$ we set the cost (d_{jk} in Eq. (1)) to capture the likelihood that an entity that checked in at c in time interval t_i will check out at c' in time interval t_j. This likelihood can be modeled using various factors.

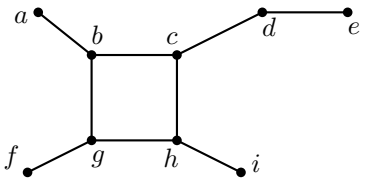


Fig. 4. A simple network with checkpoints.

We observe that also with anonymous point-based checkpoint data in a network, we can potentially reconstruct flow using the EMD. However, the number of variables needed may be large, especially if we use fine granularity of time.

4 Evaluation

This section provides an experimental validation of the use of our approach in estimating flows from checkpoint snapshots. The evaluation uses real movement data as its "ground truth", generates granulations of this data at different timestamps based on spatial tessellations as input to the EMD LP, and evaluates the accuracy of the estimates based on comparison with the original movement data. It is important to note that the experiments described in this section are not overly concerned with actually reconstructing precise moving object flows and trajectories from granular snapshots—the snapshots are too information-poor for any method to reliably achieve that. Rather, the evaluation attempts to demonstrate the extent to which our approach can capture the broader flows, directions, and distances, and show that it is flexible enough to accommodate a range of other information about constraints to movement.

Experimental setup. Our evaluations use a real data set of courier movement trajectories in central London, UK, in 2007 (the Ecourier data set[6]). The

[6] `https://en.wikipedia.org/wiki/Ecourier`

location-update frequency of trajectories varies between one coordinate every 10–30 seconds.

As discussed above, the EMD LP takes as input two snapshots of the granular distribution of spatial objects, generating a matching between the cells in one snapshot to the cells in the next snapshot as output. This matching can be directly interpreted as flow. Each input snapshot summarizes the number of objects in each cell at that time. Thus in our experiments, we spatially granulate the trajectory data by aggregating courier locations at specified times according to a raster grid of user-defined size and location. Based on preliminary studies of the data, a 22km squared area of central London was chosen for this study, and decomposed into a 40×40 raster grid for the purposes of trajectory aggregation (i.e., each cell is square with a 550m side length). Each trajectory was snapped at the relevant snapshot times to the nearest grid center, yielding a rounded ground truth that can in theory be reconstructed exactly. Our evaluation can then compare these known "ground truth" trajectories with the flows predicted by the EMD LP based only on the counts in cells.

Using the Ecourier data set ensures that our evaluation operates upon realistic movement patterns. However, the limited number of couriers in close proximity at any one time would make the task of unambiguously identifying movements in the raw courier data too simple for the EMD LP (that is, in reality, most grid cells would contain zero or one couriers at any one time). To provide a more challenging simulation of the contemporaneous movements of larger numbers of objects, we densified the trajectory data set by aggregating all courier trajectories over every day over a two month period (May–July 2007) down to a single day (that is, retaining the time-of-day portion of the trajectory time stamps, but discarding the trajectory date). Hence, our evaluation uses approximately 280 courier trajectories in our study area at any one time, ensuring that between zero and 10 couriers may appear in the same cell at a time.

Experiment 1: Flow accuracy. We begin by comparing the flows estimated by the EMD with the known "ground truth" trajectories of moving objects. Fig. 5a shows the changes in accuracy with increasing the time interval between the two input snapshots. Accuracy is measured on a per-object basis as the number of correctly estimated object movements (i.e., correct flow between an origin cell to a destination cell) divided by the total number of objects.

Broadly, Fig. 5 shows EMD estimation accuracy decreasing with increasing time interval between snapshots. This decrease is to be expected, as in longer temporal intervals, objects have a greater range of potential destinations. On average, couriers in our data set travel about 400m in 1 minute, with the fastest objects traveling 1.8km in that time ($\approx$110km/h).

Overall, the model can be said to perform relatively well. At the smallest temporal interval between snapshots (10s), the model achieves near perfect accuracy of prediction. With snapshots 2 minutes apart (120s), the model still achieves 50% accuracy in predictions.

Experiment 2: Distance and direction accuracy. The evaluations in Fig. 5 do not account for "near misses"; only estimated flows that are *exactly* correct

contribute to the accuracy or skill scores. In practice, estimations may differ in the *degree* to which they approximate the true flows. Figure 6a shows the accuracy of estimated flow *distance*, in terms of the total number of objects with estimated flows of the correct length, when compared with the total number of moving objects. As might be expected, the accuracy is moderately increased over the accuracy observed in Fig. 5.

Fig. 6b shows the accuracy of estimated flow *direction*. Averaging the direction of all flows from each cell provides an overall flow direction for that cell. The accuracy in Figure 6b is the proportion of cells with an overall estimated flow direction within 30° of the overall true flow direction. Even though the individual estimated flows might not exactly match the main flows, the response curve in Fig. 6b shows that the overall direction of estimated flows closely matches (i.e., is within 30°) of the overall direction of main flows in the majority of cases, even up to and beyond 2 minute gaps between snapshots.

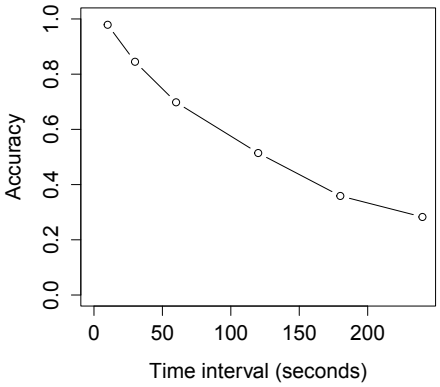
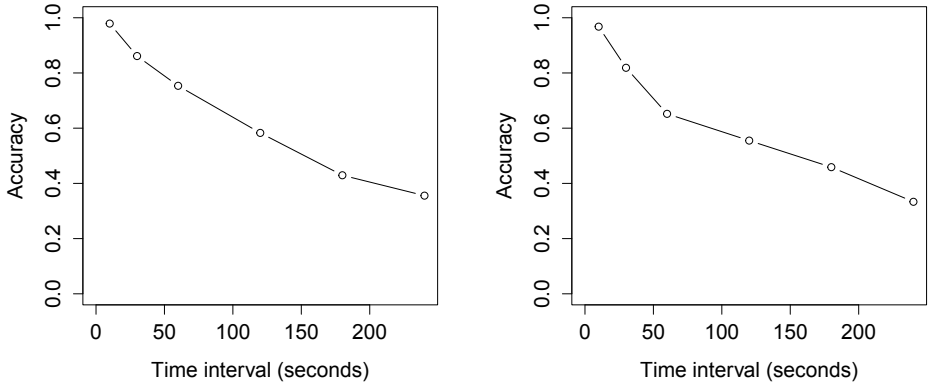


Fig. 5. Estimation accuracy of the flow.



(a) Flow-distance accuracy. (b) Cell-direction accuracy.

Fig. 6. Flow accuracy for movements with various distances.

Experiment 3: Comparison with Baseline. One further evaluation of the EMD LP flow estimations is to compare with an independent matching baseline. A natural baseline is a randomized, greedy allocation, as summarized in Algorithm 1. In short, based on the two snapshots, the algorithm randomly selects a "provider" cell with a stock of objects that must flow out. It then allocates as much of that stock as possible to the nearest "consumer" cell with a demand for

in-flowing objects. The algorithm iterates until all the stocks are exhausted and demands are satisfied.

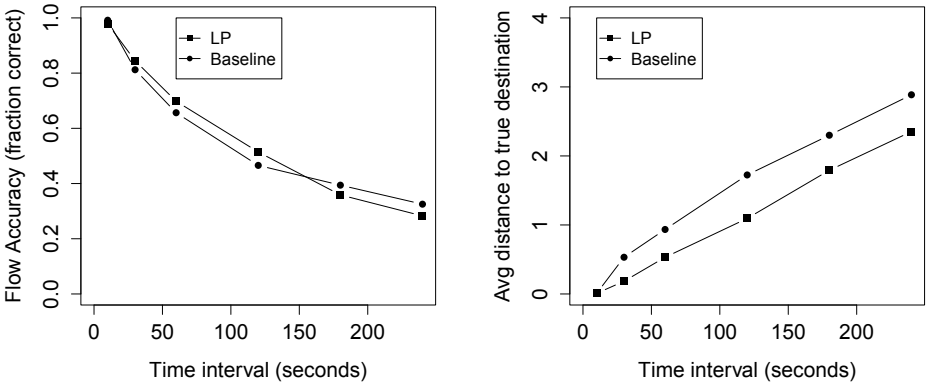
Algorithm 1: Randomized, greedy allocation Baseline

Data: Set of cells L and numbers of objects in each cell $n_s : L \to \mathbb{N}$ and $n_e : L \to \mathbb{N}$ at start and end snapshots respectively

1 Initialize the *stock* of each cell $stock : L \to \mathbb{N}$ as $stock(l) \mapsto n_s(l) - n_e(l)$;
2 Initialize $P = \{l \in L | stock(l) > 0\}$ (providers) ;
3 **while** P *is not empty* **do**
4 Select a random provider cell $p \in P$;
5 Assign as many objects as possible from $stock(p)$ to the nearest consumer cell, $c \in C$ where $stock(p) < 0$;
6 Update remaining *stock* for c and p;
7 If $stock(p) = 0$ remove p from P;

Fig. 7a compares the estimation accuracy of the EMD with the estimation accuracy of the Baseline. The response curves of the two estimations, the EMD (also shown in Fig. 5 and the Baseline, show little difference, with perhaps the EMD marginally outperforming the Baseline over shorter time intervals. However, a t-test comparing the per-cell accuracy values at each time interval revealed no significant difference between EMD and Baseline estimations (at the 95% level).

At first glance, this result is disappointing as it seems to indicate the EMD solution cannot demonstrably outperform the naïve, suboptimal Baseline. However, on closer inspection, both EMD and Baseline are fundamentally matching algorithms, using exactly the same information and constraints. Further, looking more closely at the quality of estimation, in terms of the spatial distance between estimated and true flows, does reveal a performance advantage of us-



(a) Flow estimation accuracy. (b) Target error (distance).

Fig. 7. Distance prediction accuracy.

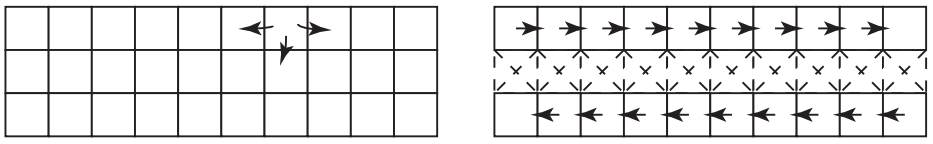
ing the EMD. Fig. 7b shows the average distance (in terms of number of cells) between the estimated target (destination cell) of flows and the true target of flows. The results show that the flows estimated by the EMD have targets that are systematically closer to the targets of the true flows than for the corresponding Baseline estimation. A t-test showed that this difference was statistically significant at the 95% level for all time intervals, except the shortest (10s).

Experiment 4: Movement constraints. One final evaluation examines the addition of movement constraints to the LP model. As discussed in Section 3, it is possible to add to the LP known constraints to movement, such as obstacles or barriers to movement. It was not possible to add these constraints to the experimental setup used in the previous experiments, because in central London at a grid size of 550m, every grid cell is effectively "connected" to every adjacent cell by at least one road. Hence, at this level of granularity, there are no obstacles to movement.

Instead, Experiment 4 "zooms in" on one road, a 16km section of the M25 London Orbital. This major motorway was frequently used by many couriers, although once again we densified the data, aggregating all the courier trips along that stretch of motorway to a single day, to ensure a sufficiently challenging, large set of contemporaneous movements. The road was then segmented into 20 1.6km long segments: 10 segments for couriers traveling east to west; 10 segments for couriers traveling west to east. Fig. 8a illustrates the cells of the granulation and their connectivity, with all neighboring cells connected. At each timestep, moving objects were assigned to cells in this granulation based on both coordinate location (provides east/west cell location) and on direction of movement, to enable disambiguation of which carriageway the object was traveling on.

Fig. 8a illustrates the cells of the granulation and their connectivity, with all neighboring cells connected. Of course, in practice we know that vehicles cannot travel in the wrong direction along a motorway carriageway, nor can they switch between carriageways directly, without first leaving the motorway and rejoining at an exit. Hence, Fig. 8b illustrates these constraints to movement, encoded through penalizing to the maximum weight disallowed movements between cells (i.e., between carriageways or in the wrong direction along a carriageway).

Fig. 9 compares the EMD estimated flows with and without the constraints to disallowed movements along the motorway. The figure shows that the EMD



(a) Fully connected road.

(b) Carriageway and direction constraints.

Fig. 8. Cells and connectivity of Experiment 4 road granulation.

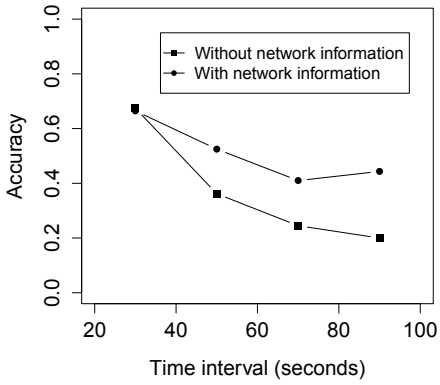


Fig. 9. Effect of movement constraints upon EMD estimation accuracy

does provide a better estimation of flows when information of underlying movement constraints are provided. As temporal granularity decreases, the difference network information makes tends to be more significant. A t-test suggested that the difference was statistically significant at the 95% level except for the 30s time intervals group.

Discussion. The four experiments described above aim to provide a picture of the strengths and weaknesses of our approach, using the EMD to reconstruct flows from granular checkpoint data. In summary, the results of these experiments indicate that the EMD:

1. is capable of regenerating flows from spatially granular checkpoint data with relatively high accuracy, certainly better than chance, especially for shorter temporal intervals where the potential for dispersion are lessened (Experiment 1);
2. is able to provide even greater reliability in generating information about broader distance and directions of flows (Experiment 2);
3. can significantly improve on the quality of estimations when compared with a naïve, suboptimal baseline matching solution, at least in terms of the spatial proximity of estimated flow targets to true flow targets; and
4. is able to incorporate information about constraints to movement, where available, and use that to improve the accuracy of estimates.

5 Conclusions and Future Work

Checkpoint data is becoming increasingly a source of data to be analyzed. This is due to both new data acquisition methods and to privacy considerations. We have shown that movement based on anonymous checkpoint data can be analyzed, and flow reconstructed, despite the low information content. We suggest the Earth Mover's Distance as a general, versatile technique to achieve this. In our

experiments we analyze the Ecourier data set and obtain meaningful results on flow, provided that the temporal resolution is relatively small. We cannot reconstruct flow if it is random, or different patterns cancel out the possibilities of detection.

The opportunities for future research abound, especially in experimentation and validation. We list several research directions of interest.

In our data set, we can expect better performance, or meaningful results over longer time periods, if we add further information like major roads. These can be incorporated using flow direction and as obstacles, as described, but also as preferred (faster) routes by lowering the distance costs between certain cells.

Intuitively, network distance is more accurate than other types of distances for network-based movement. The difficulty for applying it lies in choosing representative network nodes for cells based on which network distance can be defined. Such difficulty can be reduced by aggregating movement with a fine-grained space partition schema. Also, with point-based checkpoint data in stead of area-based one as used in this paper, network distance is naturally more suitable than Euclidean distance.

It is also interesting to analyze to what extent we can find flow patterns in other data sets, using similar approaches. These could be data sets based on mobile phone data, as in Section 3, or point-based checkpoint data in a network.

We are interested in the spatial and temporal granularities and how they affect the correctness of the flow we find. With a high spatial granularity, we will run into efficiency problems and may need to develop hierarchical methods to approximate the EMD-based flow efficiently.

We can potentially obtain better and more reliable flow when we use more than two snapshots in a single flow reconstruction. This must be modeled first, and then tested against flow reconstruction based on two snapshots only.

References

1. A. Abdul-Rahman and M. Pilouk. *Spatial Data Modelling for 3D GIS*. Springer, 2008.
2. N.V. Andrienko and G.L. Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2):205–219, 2011.
3. X. Ban, R. Herring, J.D. Margulici, and A.M. Bayen. Optimal sensor placement for freeway travel time estimation. In *Transportation and Traffic Theory 2009: Golden Jubilee (proc. ISTTT18)*, pages 697–721. Springer, 2009.
4. A. Both, M. Duckham, P. Laube, T. Wark, and J. Yeoman. Decentralized monitoring of moving objects in a transportation network augmented with checkpoints. *Comput. J.*, 56(12):1432–1449, 2013.
5. K. Buchin, B. Speckmann, and K. Verbeek. Angle-restricted steiner arborescences for flow map layout. *Algorithmica*, 72(2):656–685, 2015.
6. N.A. Giudice, L.A. Walton, and M. Worboys. The informatics of indoor and outdoor space: A research agenda. In *Proc. 2nd ACM SIGSPATIAL Int. Workshop on Indoor Spatial Awareness*, pages 47–53, 2010.

7. M. Goh. Congestion management and electronic road pricing in Singapore. *Journal of Transport Geography*, 10:29–38, 2002.
8. R.P. Greene and J.B. Pick. *Exploring the Urban Community - A GIS Approach*. Prentice Hall, 2006.
9. J. Gudmundsson, P. Laube, and T. Wolle. Movement patterns in spatio-temporal data. In S. Shekhar and H. Xiong, editors, *Encyclopedia of GIS*, pages 726–732. Springer, 2008.
10. D. Gunopulos and G. Trajcevski. Similarity in (spatial, temporal and) spatio-temporal datasets. In *Proc. 15th International Conference on Extending Database Technology, EDBT*, pages 554–557, 2012.
11. H.W. Ho, S.C. Wong, H. Yang, and B.P.Y. Loo. Cordon-based congestion pricing in a continuum traffic equilibrium system. *Transportation Research Part A: Policy and Practice*, 2005.
12. S.P. Hoogendoorn and P.H.L. Bovy. State-of-the-art of vehicular traffic flow modelling. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 215(4):283–303, 2001.
13. D. Huff. Defining and estimating a trade area. *Journal of Marketing*, 28:34–38.
14. D. Huff and W. Black. The Huff model in retrospect. *Applied Geographic Studies*, 1:83–93, 1997.
15. P. Jeszenszky and R. Weibel. Measuring boundaries in the dialect continuum. In *Proc. of AGILE*, 2015.
16. P. Laube. *Computational Movement Analysis*. Springer Briefs in Computer Science. Springer, 2014.
17. B. Mao, L. Harrie, and Y. Ban. Detection and typification of linear structures for dynamic visualization of 3D city models. *Computers, Environment and Urban Structures*, 36:233–244, 2012.
18. Tomoki Nakaya. Local spatial interaction modelling based on the geographically weighted regression approach. *GeoJournal*, 53(4):347–358, 2001.
19. T. Ott and F. Swiaczny. *Time-Integrative Geographic Information Systems*. Springer, 2001.
20. W.J. Reilly. *The Law of Retail Gravitation*. Knickerbocker Press, 1934.
21. C. Rense, S. Spaccapietra, and E. Zimányi, editors. *Mobility Data - Modelling, Management, and Understanding*. Cambridge University Press, 2013.
22. J.-P. Rodrigue, C. Comtois, and B. Slack. *The Geography of Transport Systems*. Routledge, 2006.
23. Y. Rubner, C. Tomasi, and L.J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.
24. F. Simini, M.C. Gonález, A. Maritan, and A.-L. Barabázi. A universal model for mobility and migration patterns. *Nature*, 484:96–100, 2012.
25. J. Wang, M. Duckham, and M. Worboys. A framework for models of movement in geographic space. *International Journal of Geographical Information Science*, 30, 2016. In press.
26. J. Wood. Visualizing personal progress in participatory sports cycling events. *Computer Graphics and Applications, IEEE*, 35(4):73–81, 2015.
27. J. Wood, J. Dykes, and A. Slingsby. Visualisation of origins, destinations and flows with OD maps. *The Cartographic Journal*, 47(2):117–129, 2010.
28. X. Zhang and H. Yang. The optimal cordon-based network congestion pricing problem. *Transportation Research Part B: Methodological*, 38:517–537, 2004.
29. Y. Zheng and X. Zhou, editors. *Computing with Spatial Trajectories*. Springer, 2011.