Synthesizing high-dimensional, agent-based transport demand data from two-dimensional aggregates with iterative multiple histogram matching

Author(s):
Fourie, Pieter J.

Publication Date:
2016

Permanent Link:
https://doi.org/10.3929/ethz-b-000118466

Rights / License:
In Copyright - Non-Commercial Use Permitted
Synthesizing high-dimensional, agent-based transport demand data from two-dimensional aggregates with iterative multiple histogram matching

Pieter J. Fourie

Arbeitsbericht Verkehrs- und Raumplanung 1165
January 2016
Synthesizing high-dimensional, agent-based transport demand data from two-dimensional aggregates with iterative multiple histogram matching.

Pieter J. Fourie
Future Cities Laboratory, Singapore ETH Centre, 1 CREATE Way, Singapore 138602

Introduction

Agent-based transport demand modelling requires data in highly disaggregate form. Until now, the challenge has been to generate a disaggregate description from aggregate sources, such as origin-destination matrices, combined with aggregate information on population attributes and time use, in order to produce a realistic pattern of activities and connecting trips for each agent. The process of deriving and activity-and-agent-based demand is not trivial, and its complexity and variety of data demands is frequently levelled as a criticism when comparing the agent-based approach against traditional transport demand modelling methodologies.

However, the advent of persistent position monitoring technologies (PPMTs), such as automated transit fare collection systems and cellular phones, have heralded a new era for disaggregate transport demand modelling. These ‘big data’ sources have shown themselves to be especially suited to agent-based modelling (e.g. Erath et al., 2014; McArdle et al., 2012), as they can reveal the actual demand pattern of the individual; potentially revealing the entire day tour of trips performed by each individual in a study area.

Understandably, exactly because of their high level of spatial and temporal detail, these data sources are carefully guarded and rarely make their way into the broader research community because of the associated privacy concerns. When they do, their use is usually subject to very strict data privacy controls, governed by non-disclosure agreements that consume considerable time and effort to negotiate. This makes research very difficult to replicate and validate by researchers that do not enjoy access.

Authorities and data stewards are generally happy to release aggregated totals of sensitive information; for instance, most census bureaus will release marginal totals and joint distributions of population attributes at aggregate spatial resolutions. The resolution of these aggregate distributions is generally such that no individual will be exposed, and in many cases authorities further protect individuals’ data by randomising the value of aggregates that are below a certain minimum threshold.

In the field of transportation planning, aggregate travel information might be released in the form of origin-destination matrices by mode of transport and time of day (usually morning and evening peak, and off-peak periods). Agent-based transport modellers then reverse-engineer these into activity-travel patterns for agents in a simulation.

Because of the coarseness of especially the temporal resolution, an agent-based demand that is synthesised from origin-destination matrices, will not nearly reflect the richness of the original disaggregate data source that was used to generate it. Furthermore, current approaches to representing transport demand in an aggregate form can not characterise the entire day tour of the individual; this information is lost in the process of aggregation.

There is an urgent need to capture the full potential that these new data collection technologies might offer. PPMTs have been shown to be invaluable in developing a detailed understanding of transport demand at a high temporal and spatial resolution. Furthermore, they can offer insight into the revealed time use of people within the study area, by considering the joint distribution of their full day tours of trips (assuming people mostly use the same device or smart card throughout the day). But at the same time, the privacy concerns of individuals remains of paramount importance.

This paper is an exploration of an approach that allows one to synthesise a high dimensional data set from two-dimensional aggregate distributions, of arbitrary resolution, derived from the original big data source. This high-dimensional dataset can be a reconstruction of the joint distribution of individual trips or entire
day tours of individuals within the study area. In fact, the approach shows potential to be near limitless with regard to the set of attributes whose joint distribution one attempts to reproduce.

The approach relies on Iterative histogram matching of rotations of the supplied two-dimensional aggregate distributions, in principal component and original variable space. The method makes it possible to record an entire high-dimensional dataset as a series of images, such as is shown in Figure 1, and then reconstruct a very close approximation of the original dataset from the images.

![Figure 1](Image)

**Figure 1** Two-dimensional histograms of all pairwise joint distributions of public transport trips, in principal component and original variable space, that form the primary input for high-dimensional data synthesis.

The bottom left-hand quadrant of Figure 1 shows two-dimensional histograms of all pairwise combinations of trip start / end time and origin / destination X / Y coordinates derived from public transport smart data collected during a typical weekday in Singapore. The temporal resolution is five minutes, while spatial data has been aggregated into bins of 500 metres along both the X and Y axes. For the purpose of this paper,
transit smart card transaction X / Y coordinates have been randomized to be normally distributed from their actual discrete locations at bus stops and train stations, with a standard deviation of 100 metres along both axes. The top right quadrant of Figure 1 shows the two-dimensional histograms of all pairwise combinations of principal components of the smartcard data. In all the plots, the blue end of the spectrum denotes low counts for each two-dimensional bin, while the red areas denote high count values.

In order to construct a synthetic transport demand which would very closely match the actual observed trip making recorded in the original smartcard data set, the method presented here would require the series of images as an input, along with a mapping of the colour values to count values, as well as the extent of each of the recorded variables (minimum and maximum values), and the rotation matrix that produces the principal component views.

An important first insight into the method is that it would also make it possible to construct agent-based transport demand scenarios, by simply altering any of the images that served to describe the joint distribution, using an instrument as rough as a simple digital painting program. This is possible because, in contrast to other fitting approaches such as iterative proportional fitting and generalised raking, the method of Iterative histogram matching is insensitive both to minor inconsistencies between the various two-way projections that constrains the resultant synthesised dataset; as well as the so-called zero-cell problem. It therefore presents a simple and intuitive approach to construct a high dimensional agent-based demand without the need for complex activity based demand generation.

Most importantly, this approach makes it possible to make an entire data set available to other researchers in safe and anonymous form by containing such an image series and supporting information inside a digital document.

The following section provides some background on the high dimensional data synthesis approaches in the literature, as well as the methods of principal component analysis and iterative histogram matching that were used to derive the approach presented in this document. This is followed by a description of the method and the input data that was used, as well as a section discussing the quality of results achieved. The paper ends with a discussion and an agenda for future work.

Background

Data synthesis in agent-based transport demand modeling

Data synthesis from aggregate sources is not new in the field of agent-based transportation modelling. Especially in the field of population synthesis, a lot of work has been focused on deriving the full joint distribution of population attributes from census control totals and, usually, a small anonymized micro-sample of households records that have been sanitised of any identifying information. Recently developed procedures have also looked at synthesising the actual reference sample through machine learning techniques and simulation-based approaches (e.g. Farooq et al., 2013; Sun and Erath, 2015).

Unfortunately, these techniques cannot be applied in the context of reconstructing transport demand at the level of detail that we are interested in. Firstly, a reference sample of individuals cannot be provided, as it would potentially compromise their privacy. In order to apply the population synthesis approaches therefore, one would have to synthesise a micro-sample, using for instance a Bayesian network learning approach.

However, if we want to reconstruct the entire day tour of trips at an appreciable spatial and temporal resolution, we would require a high dimensional data set with dozens of categories along each dimension. Unfortunately, Bayesian network learning requires either very limiting assumptions about the underlying distribution of variables when dealing with continuous quantities such as time and space, or a small number of categories when those dimensions are discretized.

We therefore require an alternative approach that is free from the curse of dimensionality. Multiple histogram matching was identified from the image and signal processing literature as a potential
candidate (Borgnat et al., 2012; Shapira et al., 2013), due to its capability of reshaping a ‘blob’ of data points in high-dimensional space to fit any joint distribution.

![Histogram Matching Diagram](image)

*Figure 2. Illustration of the process of histogram matching. Value $x_{\text{src}}$ is mapped to a corresponding value on the target distribution, $x_{\text{target}}$, having the same empirical cumulative probability.*

### Histogram matching

Histogram matching is a process whereby values from a source empirical cumulative distribution is transformed to those of a target distribution, such that the value chosen from the target distribution has the same cumulative probability that the original value had in its distribution. The process is illustrated in Figure 2.

Iterative histogram matching of high dimensional datasets requires one to construct marginal histograms of a large number of rotations of the target dataset. During reconstruction, an arbitrary joint distribution in high dimensional space is put through the same rotations as was recorded for the target dataset. After every rotation, the marginal histogram of the synthetic data is matched to that of the target dataset. After many rotations and repeated adjustments, the synthetic dataset is sufficiently similar to the target dataset that no point inside the synthesis will change if subjected to any further adjustments.

This process is illustrated for the joint distribution of origin X and Y coordinates of transit smart card transactions in Figure 3. In sub-figure (a), the target joint distribution 2D histogram is displayed in orange, and the initial projection of the synthetic dataset along these dimensions is shown in blue. Both data sets have already been rotated by matrix multiplication with an orthogonal matrix that is diagonal, except for the two dimensions concerned. Their marginal distributions are shown above and to the right in corresponding colours. Sub-figure (b) shows the result of histogram matching of both marginal distributions. The dataset and 2D histograms are then rotated again through an arbitrary angle, meaning their marginal distributions now appear as in sub-figure (c). Histogram matching then produces sub-figure (d), and so on.

Multiple histogram matching has been successfully employed in many signal processing applications, especially in the fields of image processing and machine vision; for instance, to adjust the dynamic range of an image to that of a target or reference image in order for image recognition to be performed under varying lighting conditions. In many cases this process is called equalisation, where the target histogram is arbitrarily taken to be a uniform distribution for all the colour channels.
In the case of higher dimensional applications to be found in the signal processing field, joint distributions appear to be relatively simple compared to those encountered in the field of transport demand modelling (e.g. Borgnat et al., 2012). Consequently, those applications only require a limited number of rotations in high-dimensional space for the synthesised data set to closely resemble the joint distribution of the target.

It is this realisation that has steered the work documented in this paper towards using two-dimensional histograms as an input, making it possible to perform rotations of the target dataset without having to record a large number of rotations and their associated marginal histograms during encoding of the target. This decision implies that constructing the marginal histograms of the rotated dataset can now only be performed in two-dimensions at a time.

Unfortunately, reproducing the pairwise joint distributions of all constituent variables does not adequately capture their joint distribution in higher dimensional space, in a similar way that performing histogram matching along the margins from a single angle for the two-dimensional distributions shown in Figure 3 does not adequately reflect the true joint structure of the orange target data set. We therefore require additional information on the joint distribution of all the variables in order to produce realistic origin destination pairs, with realistic trip start and end times. To this end, principal component analysis was selected as a means of recording the maximum amount of information on the joint distribution in the in the most parsimonious way.

Figure 3. An illustration of how data synthesis is achieved through iterative histogram matching.
**Principal component analysis (PCA)**

PCA (Jolliffe, 2002) is a very popular method of reducing the dimensionality of high dimensional data sets, and is extensively documented in the literature. For the purposes of this paper however, principal component analysis can be summarised as a way of rotating the data in high dimensional space in such a way that the marginal distribution along each of the orthogonal axes of the transformed data set has maximum variation compared to any other high dimensional rotation.

While a mathematical treatise can be found in many of the literature sources cited above, the process can be summarised in words as follows: the first principal component is a vector in high dimensional space along which variance is a maximum. The next principal component lies in the plane orthogonal to its predecessor, and this once again chosen to lie in the direction of maximum variance. By repeating the process in the plane orthogonal to each principal component vector, we reach a maximum number of orthogonal vectors equal to the number of dimensions of the original dataset. As each vector attempts to maximize variance, higher numbered principal components necessarily display less variance than their predecessors. The normalised vectors of principal components can be compiled into a rotation matrix; performing a matrix multiplication of the synthetic dataset of the same dimensions with this rotation matrix would orient the data in the same way as the principal components of the target dataset.

Recording the two-dimensional histograms of all combinations of principal components therefore carries a lot of information about the covariance of the original data. Two-dimensional histograms also make it possible to perform Iterative histogram matching along all possible two-dimensional rotations of the principal components, obviating the need to record many more rotations of the dataset along with its associated marginal distributions, as was explained above.

**Method**

The method consists of an encoding and synthesis step. Encoding is straightforward, a series of two-dimensional histograms are constructed for each pair of variables in the target dataset, given specifications of maximum resolution along each dimension. The process is repeated for its principal components.

Synthesis is more time consuming, as was alluded to in the background section. Firstly, in order to speed up convergence, an initial synthetic data set is constructed in principal component space. Then, Iterative histogram matching against multiple random rotations in both principal component and original variable space is performed, using the 2D-histograms shown in Figure 1 as target distributions.

The following subsections briefly describe the encoding and synthesis processes.

**Encoding**

1. Scale and centre the data along each dimension such that each variable has unit variance and zero mean. Record the scale and centre vectors for reconstruction.
2. Discretize each dimension to a suitable resolution, recording the range and resolution of each variable.
3. Record for each combination of now discretely-valued variables, its two-dimensional joint distribution as a 2D-histogram, as shown in the lower-left quadrant of Figure 1.
4. Perform a PCA on the original data, recording the rotation matrix for the synthesis step.
5. Discretize each principal component to a suitable resolution.
6. Record for each combination of now discretely-valued principal components, its two-dimensional joint distribution as a 2D-histogram, as in the upper-right quadrant of Figure 1.

The process can be performed for demand modelled as a series of single trips, or we can do this for the entire tour of trips, i.e. start/end time, origin/destination coordinates for each trip in the tour for persons with 1, 2, 3 or 4 trips in their daily tour (in the case of Singapore, more than 98% of transit users only have up to 4 trips recorded per day, requiring the encoding of 4 data sets of 6, 12, 18 and 24 dimensions respectively).
Synthesis

Synthesis consists of two steps: producing a synthetic sample in joint distribution space, followed by iterative histogram matching to the previously recorded two-way joint distributions.

Producing a sample of synthetic data points in joint distribution space.

The method of Iterative histogram matching ultimately produces the same result even if we only start out with a sample of points that were uniformly sampled from the n-dimensional space bounded by each variable’s range. However, the method converges much faster if we start with a relatively plausible set of points in the high dimensional space.

Sampling from the principal components and transforming back into variable space is arguably a reasonable way to produce such a plausible set of points. However, as we do not have the full joint distribution of principal components, but only the two-way joint distributions, the best way to proceed is to perform a stepwise construction. The method can be summarised as follows:

1. Start by sampling from the first two principal components to produce a set of points at least equal in number to the number of trips or tours that we intend to produce (Note that the method therefore allows us to oversample trips or tours).
2. For each point \( x = \{x_1, x_2\} \) produced in step 1, and all remaining principal components \( i \in [3..n] \), successively sample from each combination of two-way joint distributions conditional upon the last principal component value assigned to the data point, i.e. select \( x_i \) such that a uniformly sampled probability \( \text{unif}(0,1) = P(x_i | x_{i-1}) \).
3. In this way we produce a set of points that is contained within and bounded by the joint distribution of the original dataset. The process ends by transforming the data points back into the original variable space, by performing a matrix multiplication with the transpose of the principal component rotation matrix.

Synthesis through iterative histogram matching

The next step is the iterative histogram matching in original variable and principal component space, as was shown in Figure 3. The process can be summarised as follows:

1. For each pairwise combination of variables, in either original variable or principal component space, perform a matrix multiplication that will rotate both the synthetic data and the two-way joint distribution through a random angle.
2. Compiled empirical cumulative density functions (ECDFs) of the marginal distributions along both axes of the transformation, from the counts recorded in the rotated two-dimensional histogram.
3. Invert the ECDF in order for it to return a quantile for any given probability that is fed into it.
4. Compile ECDFs of the marginal distributions along both axes for the rotated synthetic data.
5. For each value in the synthetic dataset, determine its quantiles along both axes from the ECDFs.
6. For each combination of quantiles in the synthetic dataset, replace their coordinates along the rotated dimensions with the corresponding new values from the target inverted ECDFs derived in step 3.
7. Transform the synthetic data set back to its original orientation by multiplying with the transpose of the rotation matrix from step 1.
8. Repeat steps 1 to 7 for an arbitrary number of iterations or until sufficient agreement with the 2D target histogram is achieved, for instance performing a similar two-dimensional aggregation of the synthetic data and comparing cell totals with the target 2D-histogram.
9. Repeat steps 1 to 8 for each available pairwise combination of variables.
10. Transform the synthetic dataset back into principal component space (or original variable space) by performing a matrix multiplication with the PCA rotation matrix (or its transpose).
11. Repeat steps 1 to ten until sufficient agreement between all synthetic and target 2D-histograms is reached.
Sometimes one might find that some points in the synthetic dataset remain at positions outside the two-way joint distribution for many iterations. In a post-processing step, one might either remove these points from the joint distribution if one is oversampling, or force them to the nearest point in the joint distribution that has a non-zero value, using a nearest neighbour search. By repeating this post processing several times during the fitting process the number of ‘stragglers’ will decrease with increasing number of iterations.

The procedures outlined here have been implemented in R (R Core Team, 2015), with the following libraries used for ease of implementation and speed: magrittr (Bache and Wickham, 2014), dplyr (Wickham and Francois, 2015), tidy (Wickham, 2015) and data.table (Dowle et al., 2015). As this work is still exploratory and performance was not a primary concern, the R implementation was a sufficient prototype. However, future versions for higher-dimensional data sets of the process will probably be implemented in Java, with an implementation of the histogram matching speed optimization proposed by (Morovic et al., 2002).

Results

![Figure 4](image)

Figure 4 Result of fitting trip synthesized origin coordinates after 0 (a) and 50 (b) iterations of iterative histogram matching in original and principal component space. Plot (c) shows the actual joint distribution. Values are aggregated into counts within 500x500m cells, with the red end of the colour spectrum denoting relatively high counts, and blue denoting relatively low ones.

In this initial exploration, only trip-wise reconstruction was considered, in the interests of time, as higher dimensional data sets of entire tours still takes an inordinately long time to reconstruct with the prototype code. Therefore validation is done at the level of trips only.
In the procedure set out in this paper, the primary source of validation is always at hand; namely the two-way joint distributions of variables, both in principle component and original variable space. However, having secondary sources of validation helps us measure of the extent to which we have captured the true joint distribution of the target data source. Figure 4 shows one such a comparison, where the original smart card data (c) is compared with the synthesis before histogram matching was done (a) and after 50 iterations (b). Each iteration involved running matching operations for 8 random rotations of every pairwise combination of principal components and original variables, amounting to 50 x (15x8x2 + 15x8x2) = 24,000 histogram matching operations. The full sample of 3.64 million valid trips was reproduced.

Seeing as we are dealing with transportation data, the best measure of how well we have captured the full joint distribution, is to compare the distributions of travel times, trip or tour distances and travel speeds (either network routed or Euclidean speeds/distances), and the distribution of ‘activity times’, i.e. the time between trips in data sets recording the day tours of travellers.

A comparison of the distribution of speeds (based on Euclidean distance between trip origin and destination coordinates) of the synthesised dataset with that of the original smart card data, is a strong indicator of how well we capture the joint distribution of the target data set, as its calculation involves all six of the variables recorded (distance between origin and destination coordinates, divided by the difference between end time and start time). This comparison is shown in Figure 5, both before histogram matching was performed (i.e. the speed distribution produced by successive principal component sampling, shown in pink) and after 50 iterations (red line).

Figure 6 compares trip destinations departing from a selection of busy coordinates, for (a) the original smart card data and (b) the 50th iteration of the synthesised data. Colour denotes travel time, circle size denotes number of trips. The synthesis manages to capture the overall spatial distribution of trips and their associated travel times reasonably well.
Conclusion and Outlook

The approach set out in the preceding sections has been shown to produce a good approximation of the original joint distribution of public transport trips in time and space. The approach affords data stewards control over the resolution at which they provide data to the public, guaranteeing anonymity. The approach also affords a very concise way of summarising high dimensional dataset, with the possibility of
encoding the entire dataset into an image depicting its component variable pairwise joint distributions. Such an image can easily be transformed into the set of two-dimensional histograms that form the basis of the approach, if the range of each variable is known, along with the principal component rotation matrix, and the scale and centre vectors that were used to transform each variable in such a way that it has zero mean and unit variance.

It is important to note that the approach will still produce a result even when the individual pairwise combinations are not consistent; therefore it becomes possible to ‘paint’ an agent-based demand using a graphical editing tool. An existing demand might therefore be modified by painting in new origin or destination locations, or adding Euclidean trip distance as an additional variable, and forcing histogram matching against its actual or proposed future marginal distribution. Therefore this approach could be a very accessible way of drawing up scenarios for agent-based transport simulations, or for any other modelling application relying on high dimensional disaggregate data.

Initial tests with entire tours composed of up to 4 trips per person have shown promisingly good agreement with observation after a limited number of iterations. However, such higher dimensional datasets require far more time to achieve convergence along all dimensions, as the number of pairwise combinations increases dramatically with the number of variables. If we were to reconstruct the data set of all public transport users with four trips during the day, recording the origin and destination coordinates along with trip starting and end times for each trip, the current implementation would require Iterative histogram matching for each of the 276 pairwise combinations of variables, both in principal component and primary variable space. The reconstruction of higher-dimensional datasets using an improved algorithm is left to future work.

References


Erath, A., Fourie, P.J., Ordonez, S., Chakirov, 2014. Using smart card data as an Input for large scale agent-based transport simulation. Presented at the 1st International workshop on utilizing transit smart card data for service planning, Gifu University; Kyoto University; University of Miyazaki.


Wickham, H., 2015. tidyr: Easily Tidy Data with `spread()` and `gather()` Functions.