MODELING PARATRANSIT DEMAND WITH HANKEL-STRUCTURED POISSON TENSOR FACTORIZATION

Dingyi Zhuang
Department of Civil Engineering
McGill University
Montreal, Quebec, H3A 0C3, Canada
Email: dingyi.zhuang@mail.mcgill.ca
ORCID: 0000-0003-3208-6016

Lijun Sun, Corresponding Author
Department of Civil Engineering
McGill University
Montreal, Quebec, H3A 0C3, Canada
Email: lijun.sun@mcgill.ca
ORCID: 0000-0001-9488-0712

Word count: 5215 words text + 1 table(s) x 250 words (each) = 5465 words
Submission Date: August 1, 2020

ABSTRACT

As an emerging mode of public transport, paratransit plays a critical role in providing mobility services to the young, elder, and disabled persons to whom the standard public transport systems are not accessible. As paratransit is often operated in a flexible way without fixed routes and timetables, understanding passenger demand patterns becomes critical to planning and scheduling of daily operation. However, compared to other modes of transport, the demand of paratransit is in nature sparse, and thus it becomes challenging for traditional statistical models to obtain meaningful and interpretable results. The goal of this paper is to develop enhanced models to effectively learn spatiotemporal demand patterns from sparse demand data. In doing so, we introduce a model-free framework that integrates Hankel matrix transformation with probabilistic tensor factorization to enhance the interpretability of factorization models. By unveiling the spectrum that composes the original data, we discover explicit temporal patterns (e.g., weekly, monthly) from the latent modes that cannot be revealed by exiting methods. We conduct hierarchical clustering along spatial mode and discover that the separated spectrum and land use can explain the cluster formation, and even the trip purpose of paratransit. The Hankel transformation provides a powerful and effective tool to model small-scale/sparse data in a high-dimensional setting.

Keywords: Spatio-temporal traffic data, Probabilistic tensor factorization, Hankel transformation

INTRODUCTION

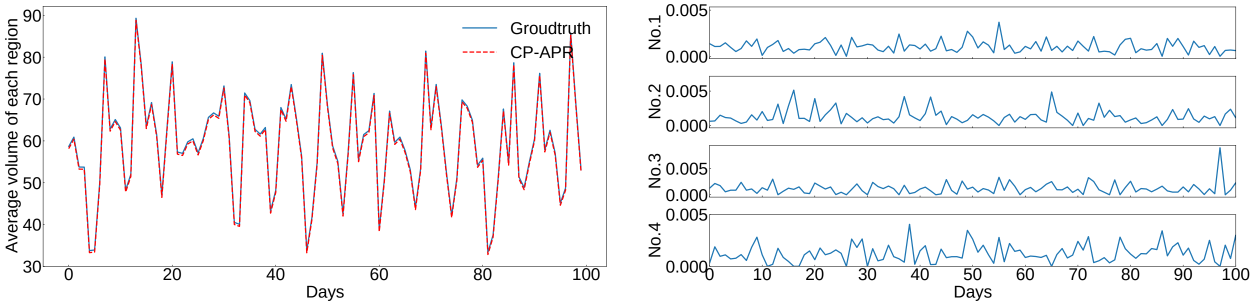
Paratransit is the service designed for the disabled people commuting by integrating public transportation and on-demand mobility modes. In general, paratransit operation data contains information of the service (e.g., origin location and boarding time). Thus, the aggregated demand can be estimated as an Origin-Time (OT) count matrix. Among real-world transportation applications, paratransit survey data have grasped less attention than other data sources like smart card data, partially because of the sparsity of paratransit demand, which makes it difficult to extract meaningful patterns. However, understanding the commuting demand of the paratransit users is also equally important to build a harmonic transportation system that covers different groups of people (*1*). Even though the trip volume of paratransit is small (with 6 million trips in 2 years in Toronto), the trip purpose and commuting routes of the paratransit should be more regular, as the disable group need health examination, working and schooling. How can we unveil hidden patterns reflecting such seasonality explicitly and efficiently from relatively small amount of data?

Tensor/matrix factorization[1] is a powerful tool for dimensional reduction by inferring principle components (i.e., factors) from incomplete high-dimensional data (*2*). In general, tensor factorization models a multi-way input tensor using low-rank latent modes where each mode can capture the hidden trends along the corresponding dimension. Generally, the tensor decomposition is defined in the real-value domain with implicit Gaussian distribution imposed, while the paratransit OT matrix is count and discrete and needs Poisson assumption on the data instead. In this aspect, probabilistic tensor factorization is more desirable since it estimates the likelihood of distribution parameters. For example, Kolda and Bader (*3*) developed alternating Poisson regression *CANDECOMP/PARAFAC* (CP-APR) factorization method for count tensor decomposition problem, which estimate the parameter $\mu_{i,t}$ for each data point $x_{i,t}$ assigned $Poisson(\mu_{i,t})$, details can be found in Section 3.4. However, as reflected in Figure 1(b), if we apply CP-APR on the paratransit OT demand matrix, hidden patterns along temporal dimension are hard to provide information of regular behavior patterns or seasonality of the disabled, even though the reconstruction result is very good in Figure 1(a). Thus, the interpretability of such probabilistic tensor factorization technique s need to improve.

In this paper, we focus on enhancing both estimation accuracy and interpretability of tensor factorization on small-scale/sparse paratransit demand data. In doing so, we develop an effective model-free framework by utilizing Hankel transformation. Considering count tensor like paratransit OT demand matrix, it can actually be regarded as multi-variate count time series. According to the theory of Fourier Transformation, it is composed of a number of bases with different frequencies. In other words, the original time series is a mixture of different regular trends, which can be effectively estimated by factorization models. However, the discrete attributes of count time series make themselves hard for convolution. Even though it is possible for count time series convolution, it can not provide dimension-based patterns and their interactions like tensor factorization does (*4*). In order to explore the hidden spectrum and bases of the trip demand matrix, we combine the idea of Hankel matrix and Poisson probabilistic tensor factorization to periodically augment the input in the probabilistic space and then decompose it. In this way, we can improve general count tensor factorization methods by exploring more interpretable patterns.

The remainder of this paper is organized as follows. Section 2 discuss works related to count tensor factorization and Hankel matrix. Section 3 introduces the probabilistic tensor factor-

[1] Because matrix is 2-dimensional tensor, we only keep term tensor factorization in the following.



(a) Average paratransit trip demand volume and CP-APR reconstruction

(b) Most significant 4 temporal patterns of CP-APR according to their λ in Equation (3)

FIGURE 1 : Paratransit OT demand matrix and its factorization results with CP-APR. The input data is the same $\mathbf{X} \in \mathbb{N}^{140\times730}$ introduced in Section 4.1 but only the first 100 days are presented. It is clear that even though (a) presents good reconstruction performance and obvious regular trends, the latent factors in (b) are still hard to understand.

ization method under Hankel transformed framework. In Section 4 we show the interpretability improvement by looking into latent modes of factorization model upon paratransit survey data. Finally we conclude and propose future tasks in Section 5.

LITERATURE REVIEW

Tensor decomposition is a powerful tool to analyze the spatio-temporal dependencies of time-varying trip record tensor (*5*). There are many excellent works that apply tensor decomposition to discover its latent patterns, smoothing the trend and even forecast (*6, 7, 8, 9*). However, most of the models focus on large-scale data sets in which it is easy to estimate principle patterns. As another important part of trip records, survey data collected from less active transportation activities, like paratransit or ambulance demand, are less studied. They are sparse, discrete and dispersed, which brings challenges for general Gaussian distribution based factorization models (*10*).

Current tensor factorization models for count data time series impose strong inductive bias. Schein et al. (*11*) imposes Poisson priors and Gamma conjugate priors to conduct variational inference to infer international dyadic events. They have follow-up works that develop non-parametric Poisson tensor factorization models and Poisson-Gamma dynamical system (*12, 13*). In developing these factorization-based models, a central challenge is to design appropriate assumptions and temporal regularization to model the dynamics and smoothness for both accuracy and smoothness (*6, 14*). Zhe and Du (*15*) use Hawkes process instead of Poisson process to model the count multi-dimensional time series. By combining Gaussian process kernel, they design new triggering functions for Hawkes and conduct variational expectation-maximization inference to capture the latent cause variables as well as the complex dependencies in tensor data. These emerging count tensor factorization methods achieve good performance in predictive tasks like imputation or prediction, but their latent components are, generally like Figure 1(b), still mixed with diverse spectrums, which does not reflect explanatory results like the seasonality of the original data.

Hankel matrix is an important concept for spectral decomposition in time series (*16*). It is widely applied in singular spectrum analysis and dynamic mode decomposition that discover physically meaningful modes in the dynamic systems, via a model-free and non-parametric way (*17, 18, 19*). Actually, the Hankel matrix-vector product of time series can solve the convolution, which can help understand the temporal behaviors of dynamic systems (*4*). Trip survey data, on the other hand, are discrete and more suitable for with Poisson assumption, which is not reasonable to directly follow their work. Thus, we still need probabilistic models to bridge the gap between discrete count and continuous data by estimating the likelihood instead of directly solving the values.

METHODOLOGY

Notation

As we receive n time-stamped demand requests for region i as the vector $\mathbf{x}_{i,1:n} = (x_{i1},\ldots,x_{in}), i = 1,\ldots,u$. Here we use the notation colon the same way in *MATLAB*, i.e. $\mathbf{x}_{i,1:n}$ refers all the n timestep (column) demand requests in region i (row). We then have a collection of entries with u unique regions under time intervals $t(=1,2,\ldots,n)$ (one day in our context). It is convenient to represent our paratransit survey data as an OT matrix, i.e. $\mathbf{X} \in \mathbb{N}^{u\times n}$. Therefore, the component $x_{i,t}$ of $\mathbf{X}$ shows the total number of paratransit demand of the i-th region at time snapshot t.

Hankel Transformation Operator

Hankel-structured tensor embeds linear time-invariant system into a high dimensional feature space, so that the data can be represented as smooth manifold in the embedded space (*20*). Now we reduce our input matrix $\mathbf{X} \in \mathbb{N}^{u\times n}$ into time series $\mathbf{x}_{1:u,1:n} = (\mathbf{x}_{1:u,1}, \mathbf{x}_{1:u,2},\ldots,\mathbf{x}_{1:u,n})$ where each element stands for all the u regional observation in certain time interval. By manually setting a window length l and $k = n-l+1$, we can define $l\times k$ Hankel matrix and Hankel Transformation Operator (HTO) $\mathscr{H}$ for time series $\mathbf{x}_{1:u,1:n}$ as

$$\mathscr{H}(\mathbf{X}) = \mathscr{H}(\mathbf{x}_{1:u,1:n}) = \begin{bmatrix} \mathbf{x}_{1:u,1} & \mathbf{x}_{1:u,2} & \cdots & \mathbf{x}_{1:u,k} \\ \mathbf{x}_{1:u,2} & \mathbf{x}_{1:u,3} & \iddots & \mathbf{x}_{1:u,k+1} \\ \vdots & \iddots & \iddots & \vdots \\ \mathbf{x}_{1:u,l} & \mathbf{x}_{1:u,l+1} & \cdots & \mathbf{x}_{1:u,n} \end{bmatrix} \tag{1}$$

To be noticed that each element in $\mathscr{H}(\mathbf{X})$ is also a vector, which means we actually obtain a 3-D tensor $\mathscr{H}(\mathbf{X}) \in \mathbb{N}^{u\times l\times k}$ after Hankel transformation. It can be found that Hankel transformation is the recursive augmentation of the original time series with the skew-diagonal elements in $\mathscr{H}(\mathbf{X})$ all the same. Therefore, the process to reverse HTO only needs to average over the skew-diagonal elements by dividing corresponding the frequency $f(t)$:

$$f(t) = \begin{cases} t, & \text{for } t = 1,\ldots,k-1, \\ k, & \text{for } t = k,\ldots,l, \\ n-t+1, & \text{for } t = l+1,\ldots,n \end{cases} \tag{2}$$

CANDECOMP/PARAFAC (CP) Factorization

In this section we briefly introduce CP decomposition, which can be regarded as high-order generalization of singular value decomposition (SVD) and principal component analysis (PCA). Even though we treat our input as matrix, HTO reshape the input with a new dimension added, which still requires tensor decomposition methods.

A N-dimensional tensor is rank one if it can be written as the outer product of N vectors. Based on that, the CP decomposition factorizes input tensor/matrix into the summation of a bunch of rank-one tensors (*2, 3*). Consider the CP factorization on $\mathscr{H}(\mathbf{X}) \in \mathbb{N}^{u\times l\times k}$:

$$\mathscr{H}(\mathbf{X}) = \sum_{r=1}^{R} \lambda_r \Theta_r^s \otimes \Theta_r^t \otimes \Theta_r^h \tag{3}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R$ represent weight parameters that indicates the importance of r-th latent component, $\Theta_r^s = (\theta_{1r}^s, \theta_{2r}^s, \ldots, \theta_{ur}^s)^T$ is the rth vector of along spatial mode. Similarly, we know the other two modes $\Theta_r^t = (\theta_{1r}^t, \theta_{2r}^t, \ldots, \theta_{lr}^t)^T$ and $\Theta_r^h = (\theta_{1r}^h, \theta_{2r}^h, \ldots, \theta_{kr}^h)^T$, representing temporal and Hankel mode respectively. The symbol "⊗" stands for vector outer product. Therefore, the elements of original tensor $\mathscr{H}(\mathbf{X})$ can be referred from decomposed tensor by

$$x_{ijk} = \sum_{r=1}^{R} \lambda_r \theta_{ir}^s \theta_{jr}^t \theta_{kr}^h \tag{4}$$

CP-APR

Even though CP factorization is widely applied in many scenarios, there is actually implicit assumption that the random variation in the tensor data follows Gaussian distribution. However, for small volume count data like paratransit surveys, it is better to describe them via a Poisson distribution (*21, 22*), i.e.

$$x_{i,t} \sim Poisson(\mu_{i,t}) \tag{5}$$

Tensor factorization based on Poisson distribution is referred as Poisson factorization. CP-APR is one of the well-performed Poisson factorization methods based on CP factorization (*3*), which is selected as the discussed model in this paper. One of the key ideas of CP-APR is to map the discrete space of $x_{i,t}$ into the continuous probabilistic space $\mu_{i,t}$, which physically means the likelihood of how many events would occur. This mapping works because the expectation of Poisson distribution fulfills $E[\mu_{i,t}] = x_{i,t}$. In this way, we impose independent Poisson distribution on each element in paratransit OT matrix and rewrite our HTO from a probabilistic aspect:

$$\mathscr{H}(\mathbf{x}_{1:u,1:n}) \sim Poisson(\begin{bmatrix} \mu_{:,1} & \mu_{:,2} & \cdots & \mu_{:,k} \\ \mu_{:,2} & \mu_{:,3} & \iddots & \mu_{:,k+1} \\ \vdots & \iddots & \iddots & \vdots \\ \mu_{:,l} & \mu_{:,l+1} & \cdots & \mu_{:,n} \end{bmatrix}) \triangleq \mathscr{H}(\mu_{:,1:n}) \tag{6}$$

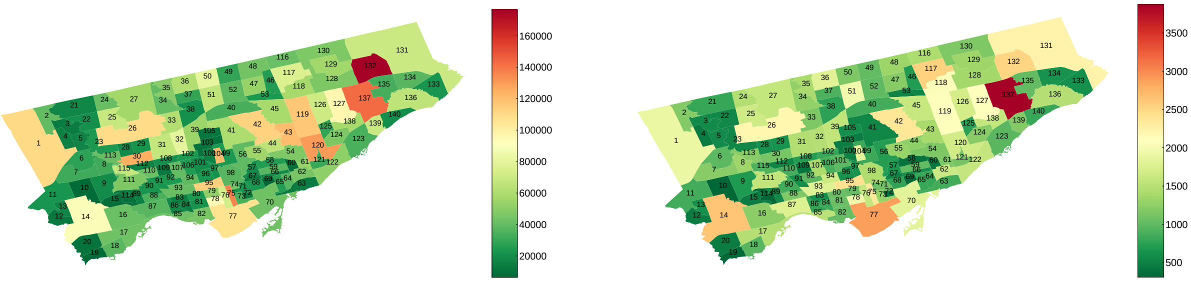
We can then conduct CP factorization on $\mathscr{H}(\mu_{:,1:n})$ to estimate the latent patterns of count tensor from the factorized modes.

NUMERICAL EXPERIMENTS

Paratransit data

As mentioned above, we use the paratransit survey data as our research dataset, which are provided by TTC's Wheel-Trans program. Wheel-Trans program provides on-demand and regular routine mobility services for disabled Toronto citizens who are unable to use conventional transit. The paratransit survey data contain 6.48 million 2-year trip records ranging from May 16th, 2017 to

May 16th, 2019, including the information of trip origins and timestamps. Because trip origins are recorded on the address level, there are over 166,446 different addresses registered. We choose the administrative region boundaries of Toronto city as our research objects, with 140 regions in total, to save memory. Selected regions can be found in Figure 2 marked with their Canadian Registration Number. Each region might contain a different number of registered addresses that the paratransit service system has received as trip origins. We delete the outlier addresses which contain trips that suddenly peak up with hundreds of records in a snapshot.



(a) Total trip volume of the selected regions

(b) Number of addresses in the selected regions

FIGURE 2 : Regional information (a) displays the summation of 2-year trip volume in each region (b) shows how many different addresses (i.e., origins of the trips) can be found inside the regions

The number of addresses is not always linearly correlated with the trip volume. As shown the regions marked red in Figure 2(a) and Figure 2(b), regions like No.132 have largest trip volumes but medium number of registered addresses in the system. This might be due to the regular usage of certain customer groups. For temporal aggregated information, we select one-day as temporal granularity. Therefore, the size of our paratransit demand OT matrix is $\mathbf{X} \in \mathbb{N}^{140\times730}$. For Hankel transformation, we manually use window length $L = 370$, slightly larger than 365 days, which is expected to cover the trends of one year (*16*). Thus our input matrix after HTO becomes $\mathscr{H}(\mathbf{X}) \in \mathbb{N}^{140\times370\times361}$. The average value of $\mathbf{X}$ is 61, meaning that each region is expected to request 61 paratransit trips every day. However, the standard deviation of $\mathbf{X}$ is also 61, which indicates that the trip demands oscillate with the same amplitude as their average. Thus, the paratransit demands are much dispersed.

Factorization Analysis

We perform CP-APR on the Hankel transformed paratransit OT matrix $\mathscr{H}(\mathbf{X}) \in \mathbb{N}^{140\times370\times361}$. Large latent component number R will give more details of the input data but will also cause overfitting upon them. Herein we choose $R = 20$ empirically to consider both the smoothing performance and the interpretability performance of CP-APR.

Temporal Mode Analysis

As presented above in Figure 1, the original paratransit trip data $\mathbf{X}$ consist of different trends which are hard to directly separate seasonality using tensor factorization. However, as shown in Figure 3, different seasonality bases are detached from the original input if we factorized upon the Hankel transformed data $\mathscr{H}(\mathbf{X})$. Similar to the idea of Fourier Transformation, the detached bases are actually the spectrum of the temporal patterns, serving as components to form the original count time series. Generally, we classify those temporally regular patterns by analyzing the intervals

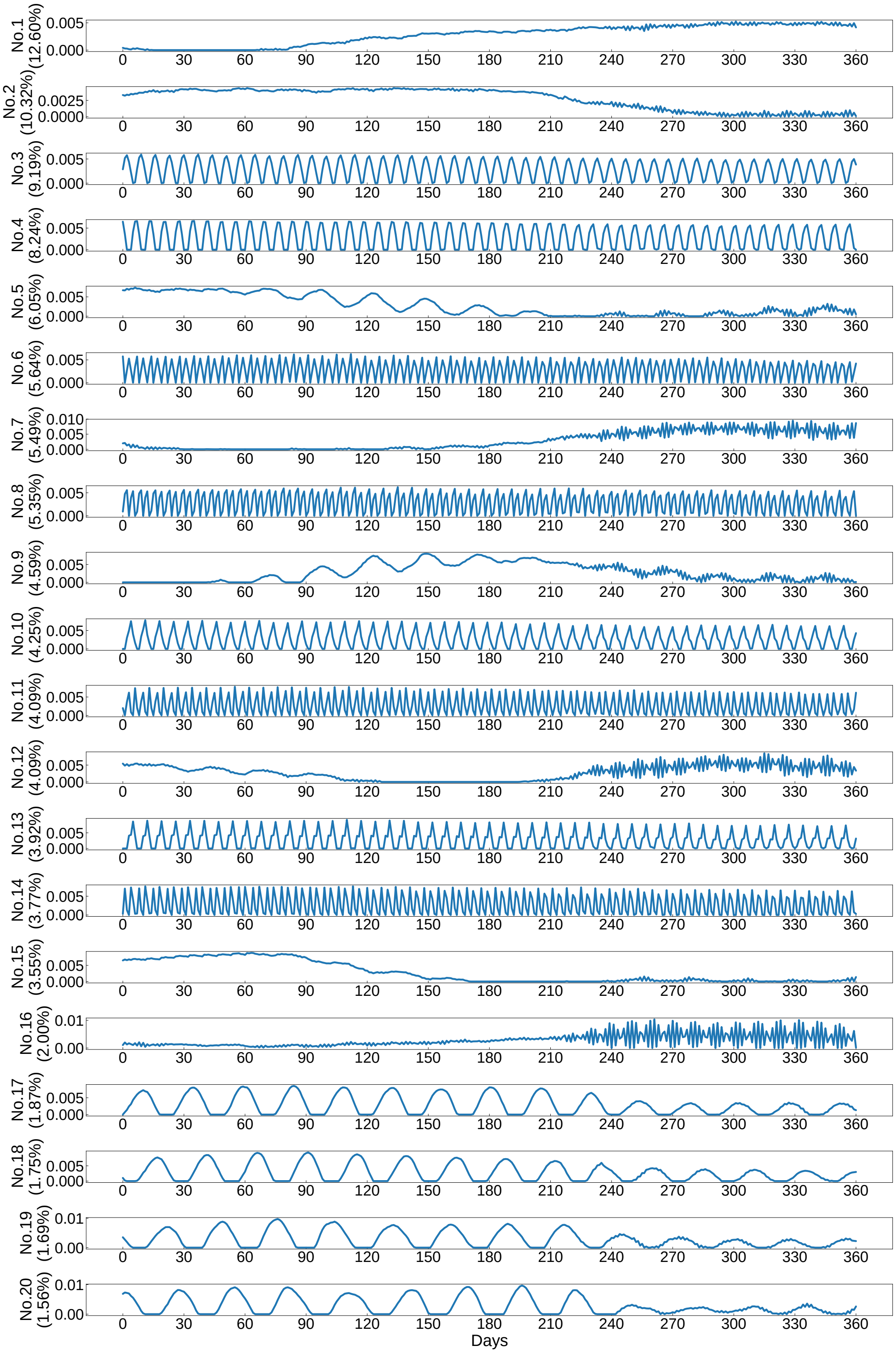


FIGURE 3 : Temporal mode Θ^t of latent factor matrices. Sorted according to their relative weight $\lambda_i / \sum_{j=1}^{20} \lambda_j \times 100\%$ from Eq (3).

between the peaks. For instance, No.3 and 4 patterns in Figure 3 have exactly the 7-day interval, which should be classified as the weekly pattern. These spectrum can help us better understand the paratransit trip behaviors.

To be more illustrative, we select some typical patterns as representatives and mark their intervals in Figure 4, where we zoomed into No.4, 6, 20, 2 and 1 patterns in Figure 3. Those patterns in Figure 4 can be divided into one-peak weekly, two-peak weekly, monthly, yearly decreasing and yearly increasing accordingly. It can be found that No.4 pattern has two-day interval with flat curve between each peak, which might infer that this trip behavior occur among weekdays because customers have to go to work or school. Whereas, No.6 pattern have two peaks where one of this peak happens exactly during the inactive interval in No.4 pattern. Thus No.6 pattern might infer the trip behavior not only contains the weekday behaviors like working, but also hang out for shopping or entertainment during the weekend. To be noticed that, interesting patterns including both No.12 and No.16, both reflect that regular behaviors (about once-two-week frequency) also exist in the monotonous patterns. These trends can be further filtered if we apply larger latent component number R. Detailed classification of typical patterns is summarized in Table 1.

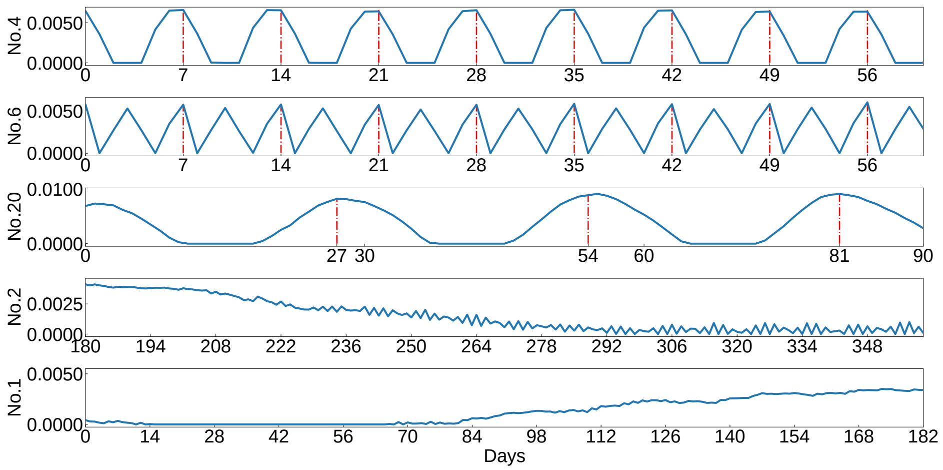


FIGURE 4 : Selected typical temporal patterns from Figure 3

Type	Frequency/days	Pattern No.
One-peak weekly	7	3,4,10,13
Two-peak weekly	7	6,8,11,14
Monthly	21-30	17,18,19,20
Yearly increase	Monotonous	1
Yearly decrease	Monotonous	2,15

TABLE 1 : Classified typical patterns in Figure 3.

Spatial Mode Clustering

Besides the factors matrix on temporal mode Θ^t, we can further explore the community structure and of the city analyze the trip purpose using the spatial mode Θ^s (*23*). We compute the conditional

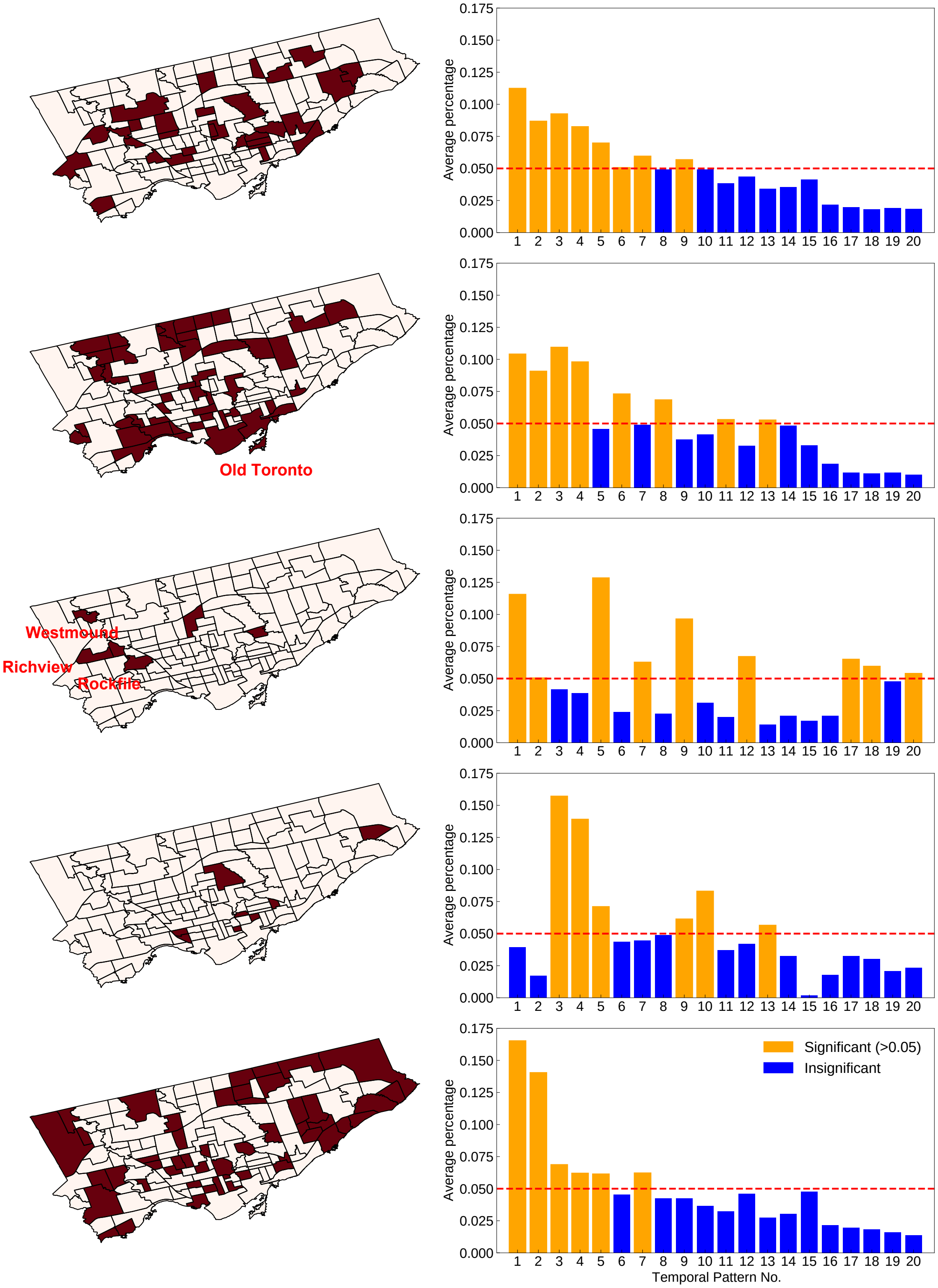


FIGURE 5 : Hierarchical clustering of conditional probability of $P(patterns|regions)$. Left: the spatial distribution of hierarchical clustering results. Right: average percentage of composition of different temporal patterns in Figure 3, with significant patterns (> 0.05) marked orange.

probability that a region belongs to a particular temporal pattern, and thus we can infer the trip behavior structure of each region from the distribution of conditional probability. For the spatial factors, given $r \in \{1, 2, \ldots, R\}$, and we can normalize this slice $\theta_{1:u,r}$ by dividing its summation to obtain the conditional probability $P(i\text{-th region}|r\text{-th component})$. Since each latent component r is associated with its weight λ_r in Equation (3). We can compute the conditional probability using Bayesian theorem as shown in Equation (7).

$$\begin{aligned} P(r\text{-th component}|i\text{-th region}) &= \frac{P(i\text{-th region}|r\text{-th component}) \times P(r\text{-th component})}{P(i\text{-th region})} \\ &\propto P(i\text{-th region}|r\text{-th component}) \times P(\lambda_r) \end{aligned} \tag{7}$$

The conditional probability $P(r\text{-th component}|i\text{-th region})$ indicates its composition of different percentage of temporal patterns. We use Agglomerative Hierarchical Clustering on the regions, where the feature of each region is $P(r\text{-th component}|i\text{-th region})$ as a vector of size $R = 20$. Hierarchical clustering is a method of cluster analysis designed to build the hierarchy of clusters (*24, 25*). Hierarchical clustering can be divided into ”Agglomerative” and ”Divisive”, where Agglomerative Hierarchical Clustering is a ”bottom-up” approach by merging pairs of clusters as one moves up the hierarchy. Specifically, each observation is considered as an individual cluster initially. At each iteration, similar cluster merge together until K clusters are formed. Therefore, the most important parameters of this method are threefold (*26*):

- The linkage criterion determines which distance to use between sets of observations. The algorithm will target at minimizing this criterion when merging pairs of clusters. For example, widely applied ”ward” linkage minimizes the variance of the clusters being merged.
- Affinity metrics used to compute the linkage, e.g. Euclidean distance or cosine similarity.
- Number of clusters to find or the linkage distance threshold above which clusters will not be merged.

Because the hierarchical clustering input $P(r\text{-th component}|i\text{-th region})$ reflects the conditional probability, the summation of each feature vector should be 1. We choose cosine similarity as the affinity metrics, use the maximum distances between all observations of the two clustered sets as linkage, and select the cluster number as 5. The parameters are tuned results according to their interpretability and the composition of different patterns. Specifically, since each feature factor is the conditional probability ranging from 0 to 1, it is better to use cosine similarity to depict the relative distance rather than absolute euclidean distance. In addition, Table 1 gives roughly 5 kinds of patterns, it is better to choose cluster number close to it.

We plot the spatially clustered results as well as their corresponding averaged distribution of latent pattern composition, as displayed in Figure 5. Because there are 20 patterns in total, significant patterns in each cluster should at least take percentage larger than $\frac{1}{20} = 0.05$. We therefore add a horizontal line with values 0.05 and mark red those patterns greater than it. From Figure 5, it can be found that spatially contiguous are generally clustered together to form communities. Meanwhile, the first 5 important temporal patterns (i.e. No.1-5) in Figure 3 appear frequently in all these clusters, indicating that they are important bases in the hierarchical clustering results. Therefore, they contain less information to distinguish clusters. We conduct empirical studies on the second and the third cluster to discuss more typical trip behaviors.

The second cluster contains No.6, 8, and 11 as significant bases, which reflects the two-peak weekly patterns. As discussed above, these kinds of patterns have strong trip demand during both weekdays and weekends, which should contains the attractions for both work and entertainment. This cluster well reflects this; for example, the Old Toronto area and The Beaches in the south contains famous Entertainment District that attracts large volume of tourism in the weekend. Large spatially contiguous regions are clustered together, demonstrating the similar land-use patterns as well as the trip demand.

The third cluster is the only cluster that includes monthly patterns, i.e. No.17-20. In addition, other significant bases, including No.7, 9 and 12 are also monotonous trend. Therefore, monthly behaviors are typical in this cluster. From Figure 2 we know that cluster contains No.7, 8 and 111 regions that locate around Richview, Westmound and Rockfile in Toronto, where large medical institution nearby like West Park Healthcare Center has large trip demand. It can be inferred that paratransit customers need to receive healthcare monthly. But weight parameter λ of these monthly patterns are less important compared to the others, which indicates that daily/weekly purpose like going to school in the second cluster takes up more quantity in the paratransit survey. Therefore, the spatial mode can serve as supplementary information for inferring the trip behaviors reflected in the temporal mode.

Hankel Mode Discussion

We also look into the details of Hankel modes Θ^h to see why Hankel transformation works. This dimension is totally manually created by stacking the fractions of time series as shown in Equation (1). As shown in Figure 6, the components generally have the same trends with temporal mode in Figure 3, but slightly different like No.5, 9, 13 and 14. Recall the direct tensor factorization upon data, shown in Figure 1, where no obvious seasonality is learned. The operation of Hankel transformation recursively augments the data to enforce the factorization model to learn regular patterns from both temporal mode Θ^t and the Hankel mode Θ^h, as mode Θ^t and Θ^h are all fractions of original time series. What's more, as mention in Section 2, Hankel matrix product with time series actually performs naive discrete convolution (*4*). In other words, $\mathscr{H}(\mathbf{X})$ already contains the spectral information of the input, while tensor factorization can then obtain hidden patterns from spectral aspect. Therefore it is found that, through simply rebuilding the input, we can enrich the interpretability of tensor factorization model.

CONCLUSION

Paratransit survey data are count data and less active than other transportation modes, which need probabilistic tensor factorization model like CP-APR to unveil their hidden patterns. However, traditional factorization model is lack of interpretation in the latent components, and fails to capture the seasonality. Hinted from the Fourier Transformation, we combine the Hankel transformation and probabilistic model to split the spectral bases of aggregated paratransit OT demand matrix and enhance the interpretability of latent components. To the best of our knowledge, we are the first to propose interpretation with Hankel transformation in the probabilistic view upon transportation datasets. We explore generally 5 types of significant patterns along temporal mode, and discover spatially contiguous communities by applying hierarchical clustering upon conditional probabil-

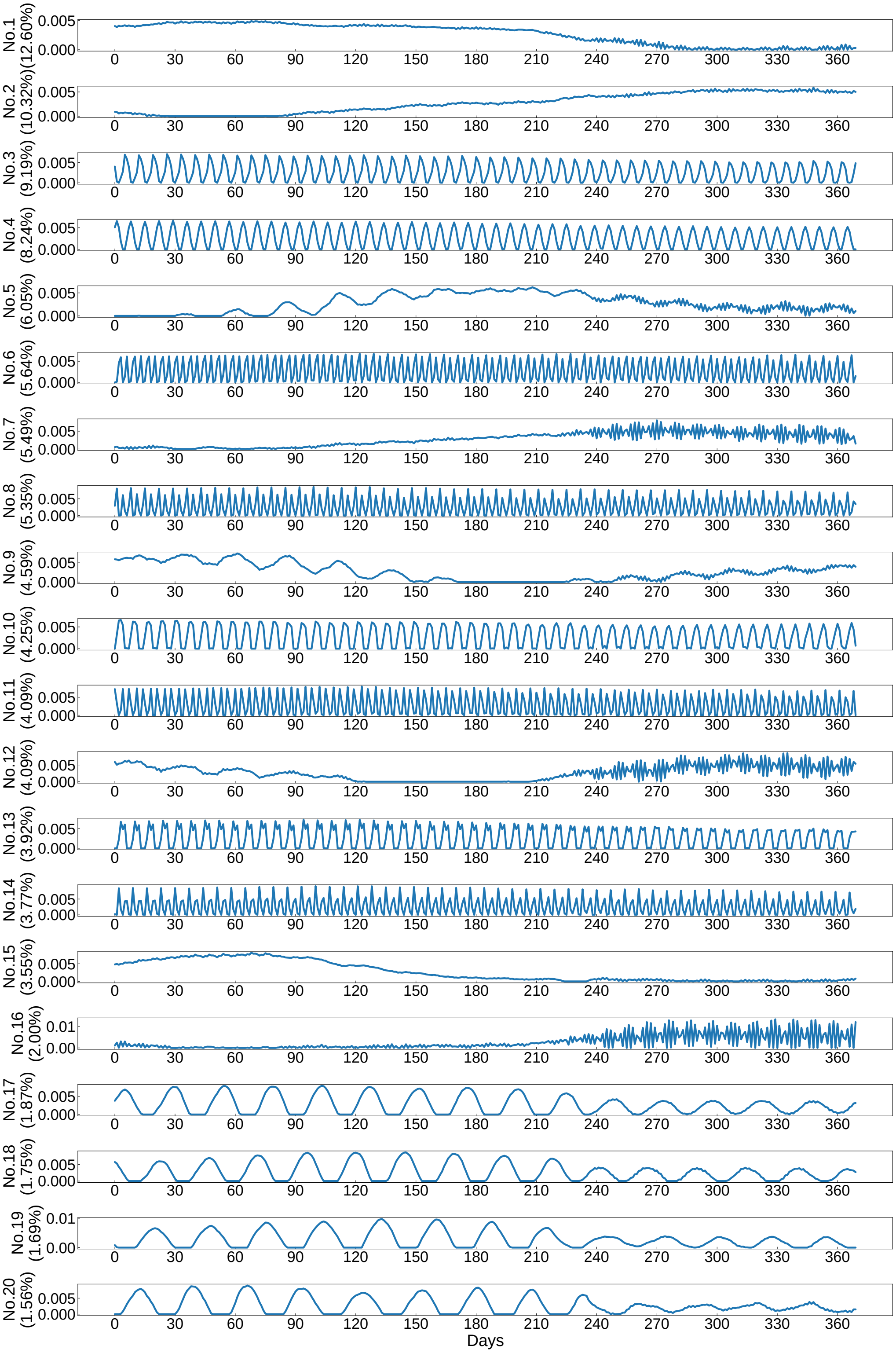


FIGURE 6 : Hankel mode Θ^h of latent factor matrices. Sorted according to their relative weight $\lambda_i / \sum_{j=1}^{20} \lambda_j \times 100\%$ from Eq (3).

ity between spatial mode and temporal mode. Given the similar patterns in Hankel mode Θ^h and temporal mode Θ^t, we discuss the data augmentation and spectral convolution effect brought by Hankel transformation. This paper only introduces the interpretability power of Hankel transformation plus probabilistic model; thereafter, we would like to demonstrate their performance on anomaly detection, imputation and prediction tasks in the future work.

ACKNOWLEDGEMENT

This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, Fonds de recherche du Quebec - Nature et technologies (FRQNT), the Institute for Data Valorisation (IVADO), and the Canada Foundation for Innovation (CFI). The authors declare no competing financial interests with respect to this work.

AUTHOR CONTRIBUTIONS

D.Z. and L.S. designed the research, performed the research, analyzed the data and wrote the paper.

REFERENCES

[1] Tuydes, H. and M. Ozen, *Scenario-Based Semidisaggregate Market Share Estimation for Proposed Paratransit System for Philippi Region, Greece*, 2009.

[2] Sun, L. and K. W. Axhausen, Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transportation Research Part B: Methodological*, Vol. 91, 2016, pp. 511–524.

[3] Kolda, T. G. and B. W. Bader, Tensor decompositions and applications. *SIAM Review*, Vol. 51, No. 3, 2009, pp. 455–500.

[4] Ju, C. and E. Solomonik, Derivation and Analysis of Fast Bilinear Algorithms for Convolution. *arXiv preprint arXiv:1910.13367*, 2019.

[5] Dunlavy, D. M., T. G. Kolda, and E. Acar, Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 5, No. 2, 2011, pp. 1–27.

[6] Sun, L. and X. Chen, Bayesian Temporal Factorization for Multidimensional Time Series Prediction. *arXiv preprint arXiv:1910.06366*, 2019.

[7] Xiong, L., X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM international conference on data mining*, SIAM, 2010, pp. 211–222.

[8] Zhou, Z., D. S. Matteson, D. B. Woodard, S. G. Henderson, and A. C. Micheas, A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association*, Vol. 110, No. 509, 2015, pp. 6–15.

[9] Cascetta, E., *Transportation systems analysis: models and applications*, Vol. 29. Springer Science & Business Media, 2009.

[10] Lee, D. D. and H. S. Seung, Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 2001, pp. 556–562.

[11] Schein, A., J. Paisley, D. M. Blei, and H. Wallach, Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1045–1054.

[12] Schein, A., M. Zhou, D. M. Blei, and H. Wallach, Bayesian Poisson Tucker decomposition for learning the structure of international relations. *arXiv preprint arXiv:1606.01855*, 2016.

[13] Schein, A., H. Wallach, and M. Zhou, Poisson-gamma dynamical systems. In *Advances in Neural Information Processing Systems*, 2016, pp. 5005–5013.

[14] Chi, E. C. and T. G. Kolda, On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, Vol. 33, No. 4, 2012, pp. 1272–1299.

[15] Zhe, S. and Y. Du, Stochastic nonparametric event-tensor decomposition. In *Advances in Neural Information Processing Systems*, 2018, pp. 6856–6866.

[16] Golyandina, N. and A. Zhigljavsky, *Singular Spectrum Analysis for time series*. Springer Science & Business Media, 2013.

[17] Hassani, H., Singular spectrum analysis: methodology and comparison, 2007.

[18] Mezić, I., Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, Vol. 41, No. 1-3, 2005, pp. 309–325.

[19] Golyandina, N. and A. Korobeynikov, Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis*, Vol. 71, 2014, pp. 934–954.

[20] Yokota, T., B. Erem, S. Guler, S. K. Warfield, and H. Hontani, Missing slice recovery for tensors using a low-rank model in embedded space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8251–8259.

[21] McCullagh, P., *Generalized linear models*. Routledge, 2018.

[22] Ridout, M., C. G. Demétrio, and J. Hinde, Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, International Biometric Society Invited Papers Cape Town, South Africa, 1998, Vol. 19, pp. 179–192.

[23] Sun, L., J. G. Jin, K. W. Axhausen, D.-H. Lee, and M. Cebrian, Quantifying long-term evolution of intra-urban spatial interactions. *Journal of The Royal Society Interface*, Vol. 12, No. 102, 2015, p. 20141089.

[24] Johnson, S. C., Hierarchical clustering schemes. *Psychometrika*, Vol. 32, No. 3, 1967, pp. 241–254.

[25] Nielsen, F., Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*, Springer, 2016, pp. 195–211.

[26] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, Vol. 12, 2011, pp. 2825–2830.