Population Location and Movement Estimation through Cross-domain Data Analysis

Xinghao Yang* and **Wei Liu**

Advanced Analytics Institute, School of Computer Science, University of Technology Sydney, Australia
Xinghao.Yang@student.uts.edu.au, Wei.Liu@uts.edu.au

Abstract

Estimations on people movement behaviour within a country can provide valuable information to government strategic resource plannings. In this paper, we propose to utilize multi-domain statistical data to estimate people movements under the assumption that most population tend to move to areas with similar or better living conditions. We design a Multi-domain Matrix Factorization (MdMF) model to discover the underlying consistency patterns from these cross-domain data and estimate the movement trends using the proposed model. This research can provide important theoretical support to government and agencies in strategic resource planning and investments.

1 Introductioin

In recent years, the rapid population growth and imbalanced population distribution are two main factors that leading to a high population pressure to Australia's major cities, which drives government consider more about: *How to estimate people's settlement patterns and movement trends?*

In this work, we take the New South Wales (NSW) state as an example, aiming to find out the population movement across different Local Government Areas (LGAs). However, the regional internal migration estimates (RIME) data only contains the number of people coming in and out of a region, without clarifying their sources/destinations. So the problem we need to address in this work can be represented in two aspects: **The Source.** Where were the intake population coming from? **The Destination.** Where were the outflow population going to? Specifically, we make the following main contributions: (1) We take the advantage of multi-domain data to evaluate the living quality, such as house adverstisement data, economy & industry and crime census data. (2) We propose a Multi-domain Matrix Factorization (MdMF) algorithm, which can simultaneously decomposes $N(\geq 3)$ domain data (see Figure 1). (3) We design a statistical learning models, which estimates the population movement trend by multiplying the consistent clustering result with RIME data.

*Contact Author

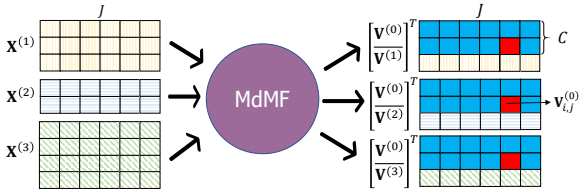

Figure 1: Multi-domain Matrix Factorization (MdMF).

2 Methodology

The cross-domain data fusion is a key technique in integrating multiple data sources to produce more consistent, accurate, and effective information. For example, [Do *et al.*, 2017] presented a collective matrix factorization method to find latent similarities of two domain data. In this paper, we design the multi-domain matrix factorization (MdMF) to match any more domain data that contains both domain-specific and domain-consistency information. Specifically, suppose we have $N(\geq 3)$ domain data $\{\mathbf{X}^{(i)} \in \Re^{D_i \times J}, i = 1, \cdots, N\}$, where D_i denotes the dimensionality of the i^{th} domain, and J is the number of LGAs, then our MdMF model is:

$$\min \mathcal{L} = \sum_{i=1}^{N} \left\| \mathbf{X}^{(i)} - \mathbf{U}^{(i)}[\mathbf{V}^{(0)}|\mathbf{V}^{(i)}]^T \right\|^2, \; s.t. \; \forall \, \mathbf{U}, \mathbf{V} \geq 0 \quad (1)$$

where each $\mathbf{X}^{(i)}$ is decomposed into a $D_i \times K$ matrix $\mathbf{U}^{(i)}$ and a $K \times J$ matrix $[\mathbf{V}^{(0)}|\mathbf{V}^{(i)}]^T$. Specifically, the common part $\mathbf{V}^{(0)}$ contains the consistency clustering result, with the element $\mathbf{V}_{i,j}^{(0)}$ (red in Figure 1) denotes the probability of the j^{th} LGA belonging to the i^{th} cluster. To achieve efficiency, we optimize each row of $\mathbf{U}$ and $\mathbf{V}$ rather than performing the optimization on the whole matrix at once. So the updating rule[1] for $\mathbf{u}_d^{(i)^T}$, $\mathbf{v}_j^{(0)}$ and $\mathbf{v}_j^{(i)}$ are:

$$\mathbf{u}_d^{(i)^T} = \mathbf{x}_{d,:}^{(i)^T} \left(\mathbf{V}^{(0i)^T} \right)^{-1} \quad (2)$$

$$\mathbf{v}_j^{(0)} = \sum_i^N \mathbf{x}_{:,j}^{(i)} \left(\sum_i^N \mathbf{U}^{(i)} \right)^{-1} \quad (3)$$

$$\mathbf{v}_j^{(i)} = \mathbf{x}_{:,j}^{(i)} \left(\mathbf{U}^{(i)} \right)^{-1} \quad (4)$$

The solution for MdMF is summarized in Algorithm 1.

[1]Derivation details can be found in our full draft: https://www.dropbox.com/s/bvil6pnxwgm5zdn/MdMF_Full_Paper.pdf?dl=0.

LGAs	SOURCE		DESTINATION	
	Top-1	Top-2	Top-1	Top-2
Blacktown	Lake Macquarie (1421)	Newcastle (1414)	Wollongong (664)	Lake Macquarie (599)
Camden	Hornsby (319)	Lake Macquarie (279)	Hawkesbury (43)	Wollondilly (43)
Campbelltown	Newcastle (592)	Lake Macquarie (590)	Wollongong (235)	Lake Macquarie (202)
Canterbury-Bankstown	Lake Macquarie (1385)	Newcastle (1329)	Lake Macquarie (721)	Newcastle (665)
Fairfield	Lake Macquarie (567)	Newcastle (540)	Lake Macquarie (266)	Newcastle (243)
Liverpool	Lake Macquarie (849)	Newcastle (843)	Wollongong (285)	Lake Macquarie (254)
Parramatta	Sydney (1013)	Hornsby (903)	Lake Macquarie (375)	Hornsby (360)
Penrith	Lake Macquarie (795)	Newcastle (783)	Wollongong (251)	Lake Macquarie (232)
The Hills Shire	Hornsby (942)	Lake Macquarie (662)	Hornsby (237)	Ku-ring-gai (222)

Table 1: The top-2 source and destination LGAs of Western Sydney area. The Numbers in brackets are the amount of population movements.

Algorithm 1 Optimization of MdMF model (1)

Input: $\mathbf{X}^{(i)}$, and ε
Output: $\mathbf{U}^{(i)}$, $\mathbf{V}^{(0)}$, $\mathbf{V}^{(i)}$
Randomly initialize all factors
Initialize $\mathcal{L}$ by a small number
repeat
  $Pre\mathcal{L} = \mathcal{L}$
  for $i = 1 : N$ **do**
    Solve $\mathbf{U}^{(i)}$ while fixing all other factor by minimizing (2)
  end for
  Solve $\mathbf{V}^{(0)}$ while fixing all other factor by minimizing (3)
  for $i = 1 : N$ **do**
    Solve $\mathbf{V}^{(i)}$ while fixing all other factor by minimizing (4)
  end for
  Compute $\mathcal{L}$ following (1)
until $\frac{Pre\mathcal{L}-\mathcal{L}}{Pre\mathcal{L}} < \varepsilon$

We then propose a statistic learning strategy to estimate the intra-state population movement. First, we proportionally subtract the amount of inter-state movement from RIME and denote the net intra-state arrival and departure vector as $\mathbf{A}^{net} = [\mathbf{A}_1^{net} \dots \mathbf{A}_J^{net}]$ and $\mathbf{D}^{net} = [\mathbf{D}_1^{net} \dots \mathbf{D}_J^{net}]$, respectively. The second step is dividing all the LGAs into two groups, i.e., the 9 Western Sydney LGAs and all the rest LGAs:

$$\mathbf{V}^{(0)} = \begin{bmatrix} \mathbf{V}^{(0),1} \\ \hline \mathbf{V}^{(0),2} \end{bmatrix}, \mathbf{A}^{net} = \begin{bmatrix} \mathbf{A}^{net,1} \\ \hline \mathbf{A}^{net,2} \end{bmatrix}, \mathbf{D}^{net} = \begin{bmatrix} \mathbf{D}^{net,1} \\ \hline \mathbf{D}^{net,2} \end{bmatrix} \quad (5)$$

Finally, the **SOURCE** and **DESTINATION** matrices are

$$\textbf{SOURCE} = \mathbf{V}^{(0),1} \cdot \left(\mathbf{V}^{(0),2} \otimes \mathbf{D}^{net,2}\right)^T \quad (6)$$

$$\textbf{DESTINATION} = \left(\mathbf{V}^{(0),1} \otimes \mathbf{D}^{net,1}\right) \cdot \mathbf{V}^{(0),2^T} \quad (7)$$

where the $\mathbf{W} \otimes \mathbf{H}$ denotes the matrix-vector product with each element of the product is $\mathbf{W}_{i,j}\mathbf{H}_j$. The **SOURCE** matrix solves the source problem - the $(i,j)^{th}$ element denotes that there were $\textbf{SOURCE}_{i,j}$ people that moved into the i^{th} Western Sydney LGA from the j^{th} other NSW LGA, and analogously for **DESTINATION**.

3 Experiments and Analysis

In this case study we focus on the year of 2015 as the RIME data depends on 2011 zoning criteria. The common clustering parameter C is empirically set to 5. We integrate house price, industry&economy and crime data to estimate living quality. For example, we select 8 features from house price data, such as, dwelling numbers, median sale price, etc. The important thing to note is that the estimated result are all local optima of our overall optimization problem which is not a strictly convex problem and can not guarantee a global optimum.

Arrival		Departure	
Sydney	20332	Sydney	21713
Lake Macquarie	11685	Hornsby	12814
Newcastle	11535	Randwick	11713
Hornsby	11461	Newcastle	11116
Ryde	9588	Lake Macquarie	10680

Table 2: The RIME data (top-5 arrival and departure LGAs).

The top-2 results of **SOURCE** and **DESTINATION** matrices are listed in Table 1. From Table 1 we can see that top-1 source areas are Lake Macquarie, Hornsby, Newcastle and Sydney. This indicates that these four LGAs are the areas with large population outflow, which is consistence with the RIME data (Table 2). The **DESTINATION** part indicates that the Lake Macquarie and Hornsby are also major migration destinations for people moving out of western Sydney. This may due to that the Lake Macquarie and Hornsby have similar living conditions with the selected LGAs. Interestingly, these four areas are geographically close to western Sydney, which makes the movement more convenient. Additionally, although Sydney ranks the top-1 in RIME arrival, it is not the best destination choice for western Sydney people. This may due to big difference of house median prices, i.e., Sydney being 72% higher than in western Sydney. The estimation results may serve as a good reference for government in making policies or for industries in making investments.

In future, we plan to query feedbacks from Sydney government about the people migration estimation results. Besides, working on quantitative analysis and comparative studies with more baseline methods should be potential research directions based on the government motivation.

References

[Do *et al.*, 2017] Quan Do, Wei Liu, and Fang Chen. Discovering both explicit and implicit similarities for cross-domain recommendation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 618–630. Springer, 2017.