Predicting the Spatio-Temporal Evolution of Chronic Diseases in Population with Human Mobility Data

Abstract

Chronic diseases like cancer and diabetes are major threats to human life. Understanding the distribution and progression of chronic disease of a population is important in assisting the allocation of treatment services and resources as well as the design of policies in preemptive healthcare. Traditional methods to obtain large scale indicators on population health, e.g., surveys and statistical analysis, can be costly and time-consuming and often lead to a coarse spatio-temporal picture. In this paper, we leverage a dataset describing the human mobility patterns of citizens in a large metropolitan area. By viewing local human lifestyles we predict the evolution rate of several chronic diseases at the level of a city neighborhood. We apply the combination of a collaborative topic modeling (CTM) and a Gaussian mixture method (GMM) to tackle the data sparsity challenge and achieve robust predictions on health conditions simultaneously. Our method enables the analysis and prediction of disease rate evolution at fine spatio-temporal scales and demonstrates the potential of incorporating datasets from mobile web sources to improve population health monitoring. Evaluations using real-world check-in and chronic disease morbidity datasets in the city of London show that the proposed CTM+GMM model outperforms various baseline methods.

1 Introduction

Recent studies show that many chronic and malignant diseases, e.g., heart disease, diabetes, and cancer, are extremely prevalent in our society [Siegel *et al.*, 2015]. To estimate disease penetration in a population and to be then able to implement appropriate health-care and preventative measures, healthcare administrators often perform statistical analysis from the records of hospital visits or conduct survey among a sample of residents. The analysis and survey usually incur high labor costs and are time-consuming. In addition, they often lead to coarse estimates both spatially and temporally.

Although chronic diseases are, to some extent, related to patients' genetics, recent studies have shown that 70% to 90% of chronic diseases can be attributed to other factors [Rappaport and Smith, 2010]. For example, there exists high causal relationship between chronic diseases and our daily habits, such as diet [McCullough *et al.*, 2002] and alcohol consumption [Martínez, 2005], which reflect different aspects of human's lifestyles. *Lifestyle* describes how people go about their everyday lives. Large scale human mobility data collected through location based social network services can act as a proxy for human lifestyle [Yuan *et al.*, 2013]. For instance, regular visits to college libraries, gyms, and lecture theaters, may correspond to the lifestyle of a college student, while a record of constant visits to meeting rooms and restaurants may indicate the lifestyle of a white-collar employee. These lifestyles are correlated with, and may reveal, certain chronic disease conditions of populations.

The connection between human mobility and diseases has been explored before. Several previous studies have investigated the outbreaks of infectious diseases via social ties and human mobility patterns [Meyers, 2007]. For chronic diseases, researchers have explored the association between human online activities and obesity [Mejova *et al.*, 2015]. However, no serious attention has previously been paid on using the lifestyles of people as the bridge between human mobility and chronic diseases nor exploring such correlation together with diseases-location similarity to predict residents' health.

In this work, we propose to seek help from the correlations between health and lifestyle reflected in human mobility to predict human's health progression of populations. Using such data for these estimates would not only give unprecedented spatial and temporal granularity to the analysis but also lower the costs of these studies and enable them to be applicable to regions for which other techniques would be deemed impractical or too expensive (e.g. developing regions). Here, we use human visitation patterns (check-ins) from Foursquare (a location-based social network) and the statistics of chronic disease morbidity in the London metropolitan area (presented on the government opening data website of the UK). We capture regional lifestyles as reflected in Foursquare mobility data, and apply a hybrid model to improve the prediction of public health conditions over simply using historic information. In summary, this paper offers the following contributions:

- We explore the correlations between human mobility patterns and health conditions and apply a method which combines Gaussian mixture models (GMM) with collab-

orative topic modeling (CTM) to predict the health levels of a population, i.e., leveraging "where they go" to help predict "how healthy they are".

- We get clues about human lifestyles from mobility patterns of residents, assuming that the groups of visited POIs are a proxy for types of lifestyle. We then exploit these inputs to identify fine-grained spatio-temporal associations between these lifestlyes and chronic diseases for local populations.
- We collect real-world chronic disease and location-based social network data to evaluate our method and analyze the correlation between lifestyles and chronic diseases. Compared with methods using historic information solely, the proposed method has a 45.7% improvement on MSE and a 1.67 times increase on R^2 score.

2 Related Work

Disease prediction. There are many studies trying to understand the spread of infectious diseases and forecast their outbreaks. Some works have analyzed the social or contact networks formed by connections among individuals and human mobility patterns to model the outbreaks of infectious diseases [Meyers, 2007], while some others utilize the large amount of users' status posts on online social networks, such as Twitter, to analyze the public health at scale [Paul and Dredze, 2011]. Some other studies target on chronic diseases. Matic et al. [Matic and Oliver, 2016] seek help from smartphone-based health applications and wearable devices to continuously record human behavior to analyze the development of human's mental health. The work [Mejova *et al.*, 2015] analyzes the data on Foursquare and Instagram and assess the relationship between fast food and obesity. Mason et al. [Mason *et al.*, 2018] even find that living further from a fast-food outlet is associated with smaller waist circumference, mostly among women. A noticeable lack of research is to analyze the effect of human mobility patterns on the development of several chronic diseases at a regional level, and also consider the similarity among these diseases and regions simultaneously. Our work aims to fill this gap.

Human mobility analysis. Patterns of human mobility are predictable and reflect how the residents of a certain area live in the physical world [Cho *et al.*, 2011]. Many studies try to learn such patterns to predict the movement of a group of people or individuals [Gao *et al.*, 2012]. Extensive research efforts have also been focused on, e.g., finding typical travel sequences with users' check-in trajectories [Zheng *et al.*, 2009], predicting users' moving patterns by exploiting both the regularity of human mobility and influence of others [Wang *et al.*, 2015], and making location recommendations with graphical models that integrates users' preferences with their sequential movement patterns [Wang *et al.*, 2016]. Different from these studies, in this paper we investigate the correlations between residents' chronic disease development and their lifestyles indicated by their frequently visited venues and mobility habits.

Topic and Gaussian mixture models. Our analysis utilizes topic modeling [Blei *et al.*, 2003] and Gaussian mixture models [Friedman and Russell, 1997]. Topic modeling has been widely used for modeling the topics of documents [Wang and Blei, 2011] and the human habits in daily lives [Yuan *et al.*, 2013]. Gaussian mixture models are often used to describe the joint effect of multiple segments and factors [Bilmes and others, 1998]. It is widely utilized on cluster prolem because of the unconstrained covariance structure and flexible application scenarios.

3 Disease Rate Evolution Prediction

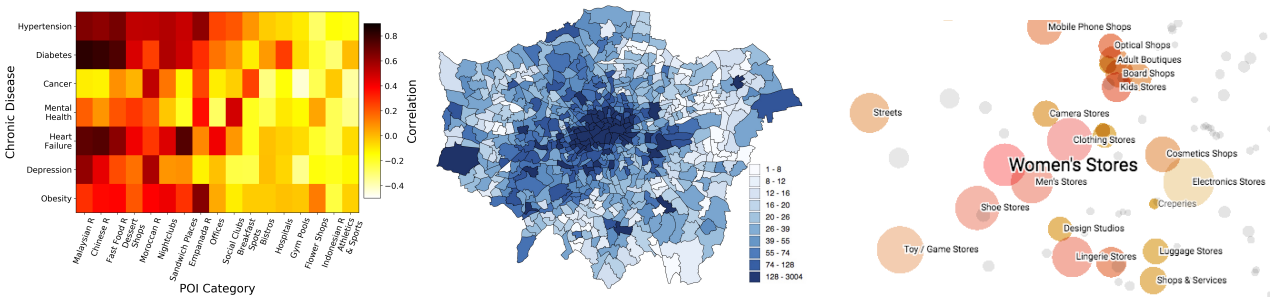
3.1 Problem Definition

Human mobility records, e.g., the check-ins, reflect people's movements in the physical world and to some extent reveal their lifestyles, which gradually affect their health conditions. Here we utilize the check-in dataset of Foursquare and the chronic disease dataset in London as an example for analysis. Foursquare dataset contains check-in records from Dec. 2010 to Dec. 2013 of 18,018 POI venues in 426 categories, e.g., fast food restaurant, gym, park, etc. There are over 4 million check-in transition pairs, in each of which we have check-in time and venue id (no user information). The chronic disease dataset contains the morbidity of 20 chronic diseases of 567 wards (self-governing units) in London area.

To exam the relationship between chronic diseases and human mobility patterns extracted from social network services, we employ Pearson correlations analysis between the evolution rate of 7 most common chronic diseases and check-in amount of 17 categories of POIs from 2010 to 2013 in London, and present the result in Figure 1(a). More specifically, for each chronic disease, we rank the disease evolution rate of 567 wards in London and split the rank list into r *segments* averagely (here we set $r = 19$, so there are 30 wards in each of the first 18 segments and 27 wards in the last one). We compute the mean value of disease evolution rate and mean check-in volume of each category of POI in every segment which forms two r-length sequences. Finally, we calculate the correlation between the two types of sequences. We can see that several categories, such as Malaysian restaurants, Chinese restaurants, and fast food restaurants, have high positive correlations with most of the 7 diseases except cancer. Some diseases have similar correlations with all the 17 categories of POIs, e.g. hypertension, heart failure, and obesity.

There are many confounding factors influencing health conditions. Our target is not to draw causal implications, but to find some of these factors to help improving prediction tasks. We plan to explore human lifestyles obtained from their footprints to POIs to help predicting their health conditions. These POIs may not all be visited by citizens living in that region, but most mobility is locally generated by definition.

People's health conditions can be treated as dynamic factors. A healthy lifestyle may not prevent illness but may reduce the risk. Therefore, we here focus on how people's lifestyles influence the development of these chronic diseases. Assume there are W regions in an area, represented as $\mathcal{R} = \{r_1, r_2, ..., r_W\}$, such as the wards in London. Let $\mathcal{D} = \{d_1, d_2, ..., d_S\}$ denote S chronic diseases and $\boldsymbol{H} \in R^{W \times S}$ denote the evolution matrix of all the S diseases in W regions , where $h_{w,s}$ is the evolution rate of disease d_s from last year to this year in region r_w (a positive number denotes an increasing rate and a negative number denotes an decreasing rate).



(a) (b) (c)

Figure 1: (a) Correlation between the evolution rate of 7 chronic diseases and check-in amount of 17 POI categories. (b) Distribution of the amount of POIs in the 630 wards of London. (c) 3-D projection of embedding results for POI categories.

However, traditional methods for chronic disease morbidity study are usually costly and time-consuming. So in practical terms, the disease data for some regions might be unavailable due to a tight budget. If we can only collect data in a subset $\widehat{R}$ regions of R, which fill in $\widehat{W}$ rows in the region-disease matrix $\boldsymbol{H}$ and let others be zero (we denote $\boldsymbol{H}$ matrix with zero rows by $\widehat{\boldsymbol{H}}$), our goal is to predict the chronic disease development of the rest $\widetilde{W} = W - \widehat{W}$ regions in set $\widetilde{\mathcal{R}}$.

3.2 Overview

We here adapt a prediction method which systematically integrates collaborative topic modeling and Gaussian mixture approach. Specifically, we firstly leverage an embedding method and a Gaussian mixture approach to aggregate the various categories of venues into several clusters according to their check-in patterns. Through that, we obtain a denser region-cluster-of-venues matrix. Then we apply a collaborative topic modeling method to extract lifestyle patterns of each region from human's check-in mobility. We hypothesize the lifestyle patterns in all regions and chronic disease evolution rates in some of the regions can be alble to to help exploring the chronic disease conditions in the missing regions.

3.3 Method

Venue Aggregation. Assuming there are N categories of venues, denoted as $\mathcal{V} = \{v_1, v_2, ...v_N\}$, e.g., French restaurant, bar, gym. We can collect the check-in amount of each category in $\mathcal{V}$ and each region in $\mathcal{R}$ (mentioned in Section 3.1), and form a check-in matrix $\boldsymbol{Y}\in R^{W\times N}$. However, the distribution of POIs is usually unbalanced. For example, Figure 1(b) shows the amount of POIs on Foursquare in the 630 wards of London during the period from 2010 to 2013. We can clearly see that central London and the region containing Heathrow Airport (the deep blue region on the left) have denser POI distribution. However, in some other regions the amount of POIs are considerably sparse, which leads to the sparsity of matrix $\boldsymbol{Y}$. To address this problem, we extract users' similar check-in preferences for some categories of POIs and aggregate these categories into a cluster. For example, if users often go shopping after having French or Italian food, we will cluster French and Italian food into a group.

To achieve the venue aggregation through check-ins, we firstly denote each category of POI as a feature vector. Those categories with similar check-in patterns will have smaller vector distances from each other. This is similar to the embedding task in language modeling, where users' transitions between those categories of POIs is analogous to a sentence, and each category of POI can be seen as a word. Here we embed each category of POI into a P-length vector through word2vec method [Rehurek and Sojka, 2010]. Figure 1(c) shows a part of the 3-D projection[1] of embedding result for all the 426 categories of POIs when P is set to 100. Each circle denotes one category. The colored circles are the categories having shortest cosine distances to the category "women's stores". Redder colors represent closer relationships. The size of a circle indicates it's "depth" from the surface of screen (foreshortening effects). We can observe that shoe stores, clothing stores, lingerie stores, and men's stores, have the most similar check-in patterns with women's stores.

We aggregate the N categories of POIs into C clusters by adopting Gaussian mixture method (GMM) [Friedman and Russell, 1997]. Compared with other cluster methods, e.g., k-means, GMM provides unconstrained covariance structure for each cluster, making the method more flexible. As illustrated in Figure 2, $\boldsymbol{\phi}\in R^{N\times P}$ is the venue-embedding matrix. The GMM can be described as follows:

$$\boldsymbol{\pi}_n \sim \text{Dirichlet}(\gamma),$$

$$\boldsymbol{\sigma}_c \sim \Gamma(\tau, \sigma_0),\ \ \boldsymbol{\mu}_c \sim \mathcal{N}(\mu_0, \nu\boldsymbol{\sigma}_c),\ \ c = 1, ..., C,$$

$$g_n \sim \text{Categorical}(\boldsymbol{\pi}_n),\ \ \boldsymbol{\phi}_n \sim \mathcal{N}(\boldsymbol{\mu}_{g_n}, \boldsymbol{\sigma}_{g_n}),\ \ n = 1, ..., N,$$

where $\boldsymbol{\pi}_n$ and g_n are the parameter of categorical distribution and the component of the n^{th} observation respectively. $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c$ are the parameters of Gaussian distribution of component c. $\gamma, \tau, \sigma_0, \mu_0, \nu$ are the shared hyperparameters. Let $\boldsymbol{X}\in R^{W\times C}$ represent the region-cluster matrix, which can be estimated through parameter $\boldsymbol{\pi}$ and check-in matrix $\boldsymbol{Y}$:

$$\boldsymbol{X} = \boldsymbol{Y} \cdot \boldsymbol{\pi}, \tag{1}$$

where $x_{i,j}$ denotes the check-in amount of venue cluster j in region r_i. Figure 2 presents the association from $\boldsymbol{Y}$ and $\boldsymbol{\pi}$ to matrix $\boldsymbol{X}$ through dotted lines.

Collaborative Topic Model. Until now, we obtain the region-cluster matrix $\boldsymbol{X}$. Our objective is to extract lifestyle information from $\boldsymbol{X}$ helping to predict chronic disease conditions in missing regions. Traditional probabilistic matrix factorization (PMF) is a perfect method for recommendation tasks, leveraging the similarity among different users and items to

[1] http://projector.tensorflow.org/

complement the user-item matrix [Salakhutdinov and Mnih, 2007]. Similarly, as mentioned in Figure 1(a), there exits similarity among diseases when analyzing the correlation between diseases and check-ins. If we assume that similarities could also be found among regions in an area, we can in the same way factorize regions' disease evolution rate matrix $\boldsymbol{H}$ into two low dimensional latent matrices $\boldsymbol{L}$ and $\boldsymbol{\Lambda}$, both with dimension K. We denote K as the amount of latent lifestyles, vector $\boldsymbol{L}_i$ as the weight of latent lifestyles in region r_i, and vector $\boldsymbol{\Lambda}_j$ as the influence from latent lifestyles on chronic disease d_j. Thus we can generate $\boldsymbol{H}$ through the distribution:

$$h_{i,j} \sim \mathcal{N}(\boldsymbol{L}_i \cdot \boldsymbol{\Lambda}_j^\top, \varsigma_{i,j}\lambda_H^2), \tag{2}$$

where $\varsigma_{i,j}$ is 0 if the data of $h_{i,j}$ is missing, and 1 otherwise. The distribution of region and disease vectors are:

$$\boldsymbol{L}_i \sim \mathcal{N}(0, \lambda_L^2 \boldsymbol{I}_K), \;\; \boldsymbol{\Lambda}_j \sim \mathcal{N}(0, \lambda_\Lambda^2 \boldsymbol{I}_K),$$

where $\boldsymbol{I}_K$ is a K-dimensional identity matrix. A common way to optimize parameters $\boldsymbol{L}$ and $\boldsymbol{\Lambda}$ is to minimize the squared-errors objective function with regularization terms:

$$\Omega = \boldsymbol{I} \odot \|\boldsymbol{H} - \boldsymbol{L}\boldsymbol{\Lambda}^\top\|_F^2 + \frac{\lambda_H^2}{\lambda_L^2}\|\boldsymbol{L}\|_F^2 + \frac{\lambda_H^2}{\lambda_\Lambda^2}\|\boldsymbol{\Lambda}\|_F^2, \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\odot$ denotes the Hadamard product operator [Kolda and Bader, 2009], i.e., element-wise multiplication.

However, we only collect the data of $\widehat{R}$ regions and lack the information to leverage similarities among the missing rows and the existing rows in matrix $\widehat{\boldsymbol{H}}$. Hence we use the region-cluster matrix $\boldsymbol{X}$ obtained in the last section. We assume each region is characterized by a particular set of lifestyles, and each cluster of categories of POIs may reflect human's various lifestyles in different probabilities. It is similar to a topic structure: if we regard all the regions as documents, the check-in patterns on various clusters of POI in a region can be considered as the words in a document. In analogy with the assumption that each topic is described by several words, each lifestyle is reflected in different check-in patterns. Intuitively, a typical topic model, LDA [Blei *et al.*, 2003], could be applied here to model the lifestyles in different regions.

However, LDA does not tackle the main problem in this work: how to leverage the lifestyles extracted from check-in mobility to fill the missing regions in matrix $\widehat{\boldsymbol{H}}$. We adopt the collaborative topic model (CTM) proposed in [Wang and Blei, 2011] here to jointly combine probabilistic matrix factorization and topic modeling. Different from the assumption in [Wang and Blei, 2011], which considers that there are offsets between the document-topic factor factorized from document-user part and document-word part respectively, we propose that people's general lifestyles, reflecting in check-in patterns and health conditions, are relatively consistent. POIs' visiting and chronic disease condition are just two views of the lifestyles, where the former one is the perspective in people's daily life activities, and the latter one is how these lifestyles influence people's health status. Therefore, we leverage the same factor $\boldsymbol{L}$, which is mentioned in PMF part, to represent region-lifestyle interactions in topic modeling.

Specifically, the generative process of the hybrid method is as follows (illustrated in the bottom part in Figure 2):

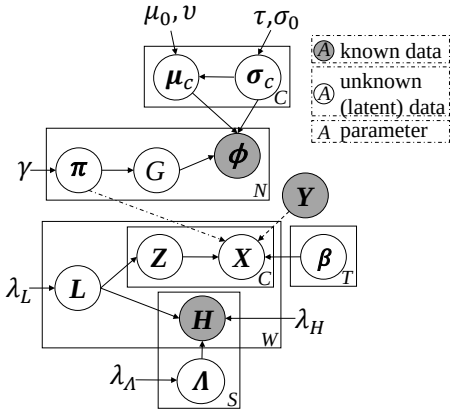


Figure 2: Illustration of CTM+GMM method.

asthma	chronic obstructive pulmonary disease (COPD)	heart failure
atrial fibrillation (AF)		learning disabilities (LD)
cancer	diabetes	mental health
chronic kidney disease (CKD)	epilepsy	obesity
	hypertension	palliative care
coronary heart disease (CHD)	hypothyroidism	smoking
dementia	heart failure due to LVD (HFLVD)	stroke or transient ischaemic attacks (stroke or TIA)
depression		

Table 1: 20 chronic diseases and their abbreviations.

1. For each chronic disease j, draw disease latent factor $\boldsymbol{\Lambda}_j \sim \mathcal{N}(0, \lambda_\Lambda^2 \boldsymbol{I})$.
2. For each region i:
   (a) Draw lifestyle factor $\boldsymbol{L}_i \sim \text{Dirichlet}(\lambda_L)$.
   (b) For each category of venues $x_{i,c}$:
      (i) Choose lifestyle assignment $z_{i,c} \sim \text{Mult}(\boldsymbol{L}_i)$.
      (ii) Choose a category of venues $x_{i,c} \sim \text{Mult}(\boldsymbol{\beta}_{z_{i,c}})$.
3. For each region-disease pair $h_{i,j}$, draw $h_{i,j} \sim \mathcal{N}(\boldsymbol{L}_i \cdot \boldsymbol{\Lambda}_j^\top, i_{i,j}\lambda_H^2)$,

where $\boldsymbol{\beta}_{z_t}$ is the distribution of POI clusters in lifestyle z_t.

Optimization. We apply EM method [Bilmes and others, 1998] to estimate the parameters $\{\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2\}(c = 1, ..., C)$ and $\boldsymbol{\pi}$ in GMM part. For the CTM part, we need to estimate the parameters $\{\boldsymbol{L}, \boldsymbol{\Lambda}\}$, where factor $\boldsymbol{L}$ is employed both in PMF and topic modeling. We iteratively optimize parameters in two steps. In the first step, for topic modeling, we estimate $\boldsymbol{L}$ through EM method, which is a typical optimization method for topic modeling. Then in the second step, we apply Gradient Descent method to estimate $\{\boldsymbol{L}, \boldsymbol{\Lambda}\}$ (we use the $\boldsymbol{L}$ factor in the first step as the initialization here). We then go back to step 1 and set $\boldsymbol{L}$ as the prior of region lifestyle distribution.

4 Evaluation

4.1 Set-Up

Data. We use three datasets in our experiment:

Foursquare dataset: The check-in data and the POI venue data of London from Dec. 2010 to Dec. 2013. The details of Foursquare dataset are introduced in Section 3.1.

Cluster 1	Cluster 2
gardens, Russian r, malls, office supplies stores, banks, art galleries, music stores, bowling alleys, theme park, dessert shops, comedy clubs, Taiwanese r, college libraries, bookstores, souvenir shops, snack places, cosmetics shops	hotels, jazz clubs, skate parks, tattoo parlors, nightlife spots, Latin American r, social clubs, cupcake shops, smoke shops, beer stores, laundry services, dinners, modern European r, convenience stores, bridges

Table 2: Several categories of POIs in 2 clusters in venue aggregation part (r: restaurants).

Boundary-line dataset of London: The dataset is collected from UK government websites[2], which contain shape file of the ward-level (electoral districts at sub-national level) boundary lines in London. In total, there are 630 wards. The shape file contains the shape line, name, and id of each ward (shape lines are shown in Figure 1(b)).

Disease dataset: We collect the data from a government open data website of UK[3]. They publish the population, the annual morbidity (value of patients/population in an area) of 19 prevalence diseases (e.g., diabetes, and obesity), and the utilization rate of one treatment method "palliative care", shown in table 1, from year 2005 to 2015. Palliative care[4] is a specialized medical care for people with life-threatening illness. Since we consider it as an indicator for the morbidity of malignant diseases, we call it "disease" here. The annual data of a year is from Apr. of one year to Mar. of the next year. For convenience, we use dataset "2013", for example, to represent the annual dataset from Apr. 2013 to Mar. 2014 in the rest of the paper. We collected the data from 2009 to 2013, consistent with the period of check-in data. The dataset contains the popularity and morbidity data of 567 wards in London (no data in the rest 63 wards).

Metrics. We apply two metrics here to evaluate the prediction performance: mean squared error (*MSE*) and R^2 score:

$$MSE(\boldsymbol{H}, \boldsymbol{H}')_{\widetilde{\mathcal{R}}} = \frac{1}{\widetilde{W} \cdot S} \sum_{r_i \in \widetilde{\mathcal{R}}} \sum_{j=1}^{S} (h_{r_i,j} - h'_{r_i,j})^2,$$

$$R^2(\boldsymbol{H}, \boldsymbol{H}')_{\widetilde{\mathcal{R}}} = \frac{1}{\widetilde{W}} \sum_{r_i \in \widetilde{\mathcal{R}}} (1 - \frac{\sum_{j=1}^{S} (h_{r_i,j} - h'_{r_i,j})^2}{\sum_{j=1}^{S} (h_{r_i,j} - \bar{h}_{r_i})^2}),$$

where $h'_{r_i,j}$ is the prediction result of item $h_{r_i,j}$ and $\bar{h}_{r_i} = \frac{1}{S}\sum_{j=1}^{S} h_{r_i,j}$. $\widetilde{W}$ is the amount of regions in testing dataset. R^2 score reflects how well the model predict values, considering the error and the mean of true values simultaneously. Here lower *MSE* and higher R^2 score represent better result.

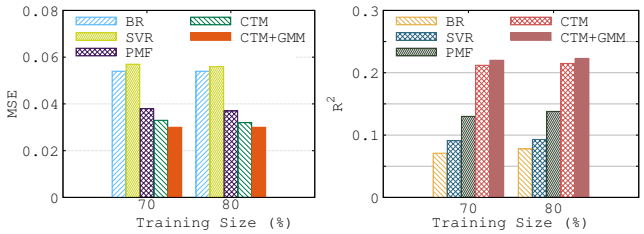
Baselines. We compare the hybrid method with 4 methods:

- *Regression*: We apply two regression methods using history data solely: boosting regression (BR) and support vector regression (SVR). They utilize the data of chronic disease evolution rate (2009 - 2012) to predict the evolution rate from 2012 to 2013.
- *Probabilistic matrix factorization* (PMF): In PMF, the region-disease matrix $\widehat{\boldsymbol{H}}$ (with missing rows) is decomposed through equation 3 ($\boldsymbol{H}$ is replaced with $\widehat{\boldsymbol{H}}$).
- *Collaborative topic modeling (CTM)*: We compare our method with CTM to emphasize the improvement through venue aggregation.

[2] https://www.ordnancesurvey.co.uk/opendatadownload/products.html#BDLINE

[3] https://data.gov.uk/dataset/quality_and_outcomes_framework_achievement_prevalence_and_exceptions_data

[4] https://en.wikipedia.org/wiki/Palliative_care



(a) *MSE* (b) R^2

Figure 3: Prediction performance of all the evaluated methods on (a) *MSE* and (b) R^2, when training size is 70% or 80%.

Disease	ER	R^2 S	Disease	ER	R^2 S
asthma	-0.03	0.155	hypertension	0.07	0.156
AF	3.17	0.166	hypothyroidism	2.94	0.129
cancer	7.12	0.150	heart failure	0.59	0.160
CKD	-2.12	0.157	HFLVD	-44.27	0.150
CHD	-1.17	0.159	LD	4.36	0.175
COPD	2.94	0.166	mental health	2.37	0.170
dementia	11.13	0.163	obesity	-16.10	0.133
depression	12.15	0.144	palliative care	21.92	0.140
diabetes	3.10	0.141	smoking	1.18	0.181
epilepsy	2.02	0.168	stroke or TIA	1.11	0.142

Table 3: Average evolution rates of the 20 diseases and their R^2 scores in prediction (**ER**: evolution rate (%), **S**: score).

4.2 Results

We evaluate and compare the prediction performance of our method with baseline models in this section. Here we embed each of the 426 categories of POIs into a 100-length vector, which forms the matrix $\phi \in R^{426\times100}$. Each vector represents POI category's features in check-in pattern space, where a shorter distance between two categories indicates that they have more similar check-in patterns. We aggregate the 426 embedded vectors into C clusters through a GMM method. Table 2 shows the categories of POIs in two typical clusters when C equals 20. We can see that most POIs in the first cluster are to some extent green (parks, gardens, etc.), artistic (art galleries, music stores, comedy clubs, etc), and reading-related (libraries and bookstores), while the second cluster mainly includes nightlife-related venues. The cluster result provides new perspectives on London citizens' visitation to POIs. Some of them are even not consistent to our experience, e.g., it is unexpected that bridges have similar check-in patterns with some clubs and nightlife spots. This may be because the fantastic lighting systems of some famous bridges in London, such as Albert Bridge and Tower Bridge, attract a large number of viewers during the night hours.

Next, we predict the missing rows in matrix $\boldsymbol{H}$, in which

item $h_{i,j}$ is the evolution rate of disease d_j in ward r_i from year 2012 to 2013. Firstly, we obtain the region-cluster matrix $\boldsymbol{X}$ through Equation 1. Then, the latent lifestyle factor $\boldsymbol{L}$ and latent chronic disease factor $\boldsymbol{\Lambda}$ are estimated through CTM in Section 3.3. Finally, we predict the missing regions $\widetilde{\mathcal{R}}$ in matrix $\boldsymbol{H}$ through Equation 2. We randomly select 70% and 80% rows in matrix $\boldsymbol{H}$ as training data and evaluate the prediction performance in the rest rows. Due to the space limit, for all the methods, we tune the parameters through the training set and show the best prediction results in testing set. We test each method 20 times and show the average results in Figure 3. The latent dimension length K and the amount of lifestyles T (in CTM part) are set to 50, and C, the amount of component (in GMM part), is set to 20. It is obvious that CTM+GMM has the best performance on the two metrics. The two regression methods have the highest *MSE* results and lowest R^2 scores. We can infer that the historic evolutions of these chronic diseases cannot provide enough regular patterns (e.g. linear or periodic patterns) for future prediction. Compared with traditional PMF, CTM has an average improvement of 13% in *MSE* and 59% in R^2 score. Moreover, when we add the GMM part to CTM, it furtherly achieves a 7.6% improvement in *MSE* and a 3.7% improvement in R^2 score.

We show the average evolution rate of all the diseases and their R^2 scores in predictions through CTM+GMM method from 2012 to 2013 in Table 3 (different with the definition in Section 4.1, we here show the R^2 scores along each disease). The Pearson correlation between evolution rate list and R^2 score list in the table is 0.111, which means that there is no obvious correlation between these two factors: the predictability of a disease is neither positively nor negatively related to its own evolution rate. In this table, we can see the significant growth of palliative care during that year, indicating that the morbidity of malignant diseases, e.g., cancer, had increased. The morbidity of all the psychiatric and mental diseases, e.g. depression, dementia, and mental health, dramatically increased in that period. Actually, as illustrated in [Muliyala and Varghese, 2010], depression and dementia have complex relationships: depression has been both a risk factor and a prodrome of dementia, which is also reflected in the close evolution rates between them. Good news is that the obesity cases in London decreased in the same period, which has been a serious health concern in UK for a long time. Moreover, we observe that obesity has a relatively low prediction R^2 score in the experiment through both our method and baselines. This may be because of that different with some incurable chronic diseases, e.g., diabetes [Etuk and others, 2010], of which the morbidity is more stable, people can reduce their weight through various approaches, making the evolution of obesity more difficult to predict.

5 Lifestyle and Chronic Diseases

We present the correlation between the 20 chronic diseases and various lifestyles. We leverage the check-in pattern as a projector of lifestyle. Specifically, we use the lifestyle-cluster factor ($\boldsymbol{\beta}$ in Section 3.3), and the disease-lifestyle factor $\boldsymbol{\Lambda}$, to uncover the hidden relationship between chronic diseases and POI venues. Some of the observed correlations are consistent with the research findings in clinical medicine and physiology, while some others provide new insight on some open problems. We list 3 typical check-in lifestyles and the highly correlated diseases in Table 4.

Check-in Lifestyles	**Correlated Dis**
government buildings, pubs, sushi r, plazas, candy stores, steakhouses, burger joints, Subways, Chinese r, churches, bookstores, fast food r, coffee shops, nightclubs	heart failure, attacks, asthma, COPD, coronary heart disease, diabetes, hypertension, obesity, HFLVD, stroke or TIA
Indian r, convention centers, fast food r, sandwich places, grocery stores, Thai r, dim sum r, hotel bars, bakeries, hostels, nightclubs, fried chicken joints	
dessert shops, organic groceries, gay bars, middle eastern r, sports bars, cocktail bars, whisky bars, nightclubs, offices, boutiques, Chinese r, hotel bars, hotels, Italian r	cancer, smoking chronic kidney, depression, disease, obesity,

Table 4: Typical check-in lifestyles and the highly-correlated diseases (Dis: Diseases, r: restaurants).

We can observe that the top two lifestyles imply a high demand of fast food: fast food restaurants, pizza places, fried chicken joints, and the high demand of Asian food: Chinese restaurants, sushi restaurants, etc. These two lifestyles show close relevance with 10 chronic diseases listed on the right. This is consultant with our empirical insight that some diseases, like heart-related diseases, hypertension, obesity, and diabetes, are correlated with diets that are too high in sugar and fat (Asian foods are more rich of ingredients). Additionally, asthma and chronic obstructive pulmonary disease are also highly correlated with these two lifestyles. This is consistent with previous studies in physiology [Rosenkranz *et al.*, 2010] claiming that a high-fat diet may contribute to chronic inflammatory diseases of the airway and lungs.

The third lifestyle is alcohol-oriented, where 6 of the 14 categories of POIs are bars or clubs. The induction of alcohol in some diseases, like cancer, depression, have been proved in previous studies [Martínez, 2005]. However, as illustrated in [White *et al.*, 2009], evidence for an association between alcohol consumption and chronic kidney disease is conflicting. Here we obtain the inspiration from large-scale human mobility data and provide an intuitive perspective on the positive correlation between lifestyles surrounded by alcohol and chronic kidney disease.

6 Conclusion

In this paper, we leveraged human mobility data to help studying the evolution of human health conditions. We embed the POI venues into vectors according to human transition patterns to capture their semantic meanings. Then we combined Gaussian mixture methods with collaborative topic modeling to predict health conditions, which dealt with data sparsity and extracted human lifestyles from check-in patterns. In particular, we studied the correlation between lifestyles and the development of chronic diseases from a new perspective. Extensive experiments using real-world datasets indicated that CTM+GMM had a significant improvement on prediction tasks compared to other methods.

References

[Bilmes and others, 1998] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *ICSI*, 4(510):126, 1998.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[Cho *et al.*, 2011] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. Friendship and mobility: User movement in location-based social networks. In *SIGKDD*, pages 1082–1090, 2011.

[Etuk and others, 2010] EU Etuk et al. Animals models for studying diabetes mellitus. *Agric Biol JN Am*, 1(2):130–134, 2010.

[Friedman and Russell, 1997] Nir Friedman and Stuart Russell. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the Thirteenth conference on UAI*, pages 175–181. Morgan Kaufmann Publishers Inc., 1997.

[Gao *et al.*, 2012] Huiji Gao, Jiliang Tang, and Huan Liu. gscorr: Modeling geo-social correlations for new checkins on location-based social networks. In *CIKM*, pages 1582–1586, 2012.

[Kolda and Bader, 2009] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[Martínez, 2005] María Elena Martínez. Primary prevention of colorectal cancer: lifestyle, nutrition, exercise. In *Tumor Prevention and Genetics III*, pages 177–211. Springer, 2005.

[Mason *et al.*, 2018] Kate E Mason, Neil Pearce, and Steven Cummins. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from uk biobank. *The Lancet Public Health*, 3(1):e24–e33, 2018.

[Matic and Oliver, 2016] Aleksandar Matic and Nuria Oliver. The untapped opportunity of mobile network data for mental health. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 285–288, 2016.

[McCullough *et al.*, 2002] Marjorie L McCullough, Diane Feskanich, Meir J Stampfer, Edward L Giovannucci, Eric B Rimm, Frank B Hu, Donna Spiegelman, David J Hunter, Graham A Colditz, and Walter C Willett. Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance. *The American journal of clinical nutrition*, 76(6):1261–1271, 2002.

[Mejova *et al.*, 2015] Yelena Mejova, Hamed Haddadi, Anastasios Noulas, and Ingmar Weber. # foodporn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health 2015*, pages 51–58. ACM, 2015.

[Meyers, 2007] Lauren Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44(1):63–86, 2007.

[Muliyala and Varghese, 2010] Krishna Prasad Muliyala and Mathew Varghese. The complex relationship between depression and dementia. *Annals of Indian Academy of Neurology*, 13(Suppl2):S69, 2010.

[Paul and Dredze, 2011] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272, 2011.

[Rappaport and Smith, 2010] Stephen M Rappaport and Martyn T Smith. Environment and disease risks. *Science*, 330(6003):460–461, 2010.

[Rehurek and Sojka, 2010] Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, May 2010.

[Rosenkranz *et al.*, 2010] Sara K Rosenkranz, Dana K Townsend, Suzanne E Steffens, and Craig A Harms. Effects of a high-fat meal on pulmonary function in healthy subjects. *European journal of applied physiology*, 109(3):499–506, 2010.

[Salakhutdinov and Mnih, 2007] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Nips*, volume 1, pages 2–1, 2007.

[Siegel *et al.*, 2015] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29, 2015.

[Wang and Blei, 2011] Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *SIGKDD*, pages 448–456. ACM, 2011.

[Wang *et al.*, 2015] Yingzi Wang, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, and Yong Rui. Regularity and conformity: Location prediction using heterogeneous mobility data. In *SIGKDD*, pages 1275–1284, 2015.

[Wang *et al.*, 2016] Weiqing Wang, Hongzhi Yin, Shazia Wasim Sadiq, Ling Chen, Min Xie, and Xiaofang Zhou. SPORE: A sequential personalized spatial item recommender system. In *ICDE*, pages 954–965, 2016.

[White *et al.*, 2009] Sarah L White, Kevan R Polkinghorne, Alan Cass, Jonathan E Shaw, Robert C Atkins, and Steven J Chadban. Alcohol consumption and 5-year onset of chronic kidney disease: the ausdiab study. *Nephrology Dialysis Transplantation*, 24(8):2464–2472, 2009.

[Yuan *et al.*, 2013] Nicholas Jing Yuan, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie. We know how you live: exploring the spectrum of urban lifestyles. In *COSN*, pages 3–14. ACM, 2013.

[Zheng *et al.*, 2009] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, pages 791–800, 2009.