

DOCTORAL THESIS

Probabilistic Models for Spatially Aggregated Data

Yusuke Tanaka

*A dissertation submitted to the Department of Systems Science
in partial fulfillment of the requirements for the degree of*

Doctor of Informatics

at the

Kyoto University

Supervisor: Toshiyuki Tanaka

March 2020

Abstract

Location data have played an important role in improving the life quality of people in various aspects such as socio-economics, epidemiology, public health, public security, urban planning, navigation systems, and location-based marketing. Due to practical reasons (e.g., privacy concerns), location data at the individual-level are often *aggregated* over spatial regions. Data aggregation is an essential process; however, it makes it difficult to apply existing methods for spatial data analysis, most of which assume that individual-level data are available. One thus requires a framework for handling spatially aggregated data, which sheds light on the more effective utilization of aggregated data to solve the problems involved in human life and city environments.

In this thesis, we describe novel computational tools that support prediction and knowledge discovery from aggregated data. Our solution is a model that integrates an aggregation process, a critical functionality for handling data aggregated over regions; this allows for effective model learning based on aggregated data. We specifically propose probabilistic models and their inference algorithms that address the following two problems: 1) refining coarse-grained aggregated data; 2) inferring latent people flow from aggregated population data. We demonstrate the effectiveness of the proposed models by extensive experiments on real-world data sets.

Acknowledgements

I am deeply grateful to my advisor, Prof. Toshiyuki Tanaka, for his insightful advice and considerate support. This thesis would not have been possible without his support. I would also like to thank my thesis committee members, Prof. Shin Ishii and Prof. Hidetoshi Shimodaira, who have provided valuable and constructive comments.

The research for this thesis was conducted at NTT Service Evolution Laboratories. I would like to thank all of the researchers who have collaborated with me. I am especially grateful to my mentor, Dr. Takeshi Kurashima. He taught me how to identify and work on research questions that matter, which is an invaluable skill for a researcher. I would like to thank Dr. Hiroyuki Toda, leader of our group, for his great instruction and broad support. I would also like to thank Dr. Hiroshi Sawada, former leader of our group, and Dr. Yasuhiro Fujiwara, my supervisor at that time, for providing valuable advice and support.

I would like to thank my great collaborators in NTT Communication Science Laboratories. I learned a lot from Dr. Tomoharu Iwata, NTT distinguished researcher. His advice on research is invaluable and has always inspired me. I would like to thank Dr. Naonori Ueda, NTT Fellow, who was also my master's degree advisor at Kyoto University, for valuable instruction and constant encouragement. He taught me the fundamental concepts and techniques of machine learning.

I would like to thank Emer. Prof. Shigeo Abe, who was my bachelor's degree advisor at Kobe University, for teaching me the fundamentals of pattern recognition and the basics of research methods.

Finally, I am very grateful to my family for their love and support.

Contents

Abstract	i
Acknowledgements	ii
Contents	iii
List of Figures	vi
List of Tables	vii
1 Introduction	1
1.1 Motivation	1
1.1.1 Data and applications	1
1.1.2 Challenges	2
1.2 Thesis overview and contributions	3
1.2.1 Overview	3
1.2.2 Part 1 – Refining coarse-grained aggregated data	4
1.2.3 Part 2 – Inferring latent people flow from aggregated population data	6
1.3 Mathematical notations	7
I Refining Coarse-grained Aggregated Data	8
2 Introduction and survey	10
2.1 Introduction	10
2.2 Related work	12
2.2.1 Categorization	12
2.2.2 Regression approach	12
2.2.3 Multivariate approach	13
2.3 Problem formulation	14
2.4 Roadmap	15
3 Spatially Aggregated Gaussian Processes with a Single Output	17

3.1	Overview	17
3.2	Model	17
3.3	Inference	19
3.3.1	Outline	19
3.3.2	Marginal likelihood	19
3.3.3	Posterior GP	21
3.3.4	Approximation of the integral over regions	23
4	Regression approach	24
4.1	Overview	24
4.2	Proposed model	25
4.2.1	Model	25
4.2.2	Inference	29
4.3	Experiments	32
4.3.1	Data	32
4.3.2	Refinement task	32
4.3.3	Setup of 2-stage SAGP model	34
4.3.4	Baselines	34
4.3.5	Results	36
4.4	Discussion	41
5	Multivariate approach	43
5.1	Overview	43
5.2	Proposed model	44
5.2.1	Model	44
5.2.2	Inference	49
5.3	Experiments	51
5.3.1	Experimental setup	51
5.3.2	Setup of SAGP-M	52
5.3.3	Baselines	52
5.3.4	Results for the case of a single city	53
5.3.5	Results for the case of two cities	58
5.4	Discussion	59
II	Inferring Latent People Flow from Aggregated Population Data	60
6	Inferring Latent People Flow from Aggregated Data at Limited Locations	62
6.1	Introduction	62
6.2	Related work	64

6.2.1	Collective graphical models	64
6.2.2	Information diffusion models	65
6.2.3	Urban computing	66
6.3	Problem formulation	67
6.4	Proposed model	67
6.4.1	Model	67
6.4.2	Inference	71
6.5	Experiments	74
6.5.1	Data	74
6.5.2	Evaluation metric	76
6.5.3	Setup of T-CFDM	76
6.5.4	Baselines	76
6.5.5	Results	77
6.6	Discussion	83
III	Conclusion and Future Directions	84
7	Conclusion	86
7.1	Summary of contributions	86
7.2	Future directions	88
A	On integrability of the observation model in SAGP-S	90
	Bibliography	91
	List of Publications	100

List of Figures

1.1	Data illustration.	2
2.1	The spatial distribution of poverty in New York City at different spatial granularities.	11
4.1	Generative process for 2-stage SAGP model.	26
4.2	Comparison of the refined poverty rate data in New York City (regression approaches).	38
4.3	Comparison of the refined recycle diversion rate data in New York City (regression approaches).	38
4.4	Comparison of the refined poverty rate data in Chicago (regression approaches).	39
4.5	Visualization of the prediction variances for auxiliary data sets in New York City.	42
5.1	Generative process for SAGP-M.	45
5.2	Graphical model representation of SAGP-M.	48
5.3	Comparison of the refined poverty rate data in New York City.	55
5.4	Comparison of the refined recycle diversion rate data in New York City.	55
5.5	Visualization of the estimated parameters.	57
5.6	Comparison of the refined poverty rate data in Chicago (transfer learning across two cities).	58
6.1	The problem we focus on: Inferring latent people flows between locations from aggregated population data.	63
6.2	Visualization of travel duration probabilities.	71
6.3	Visualization of transition matrices and travel duration probabilities.	81
6.4	Comparison of pedestrian flows between booths in exhibition halls.	82

List of Tables

1.1	Examples of location point and value for each kind of individual-level data set.	2
2.1	Notation for spatially aggregated data.	15
4.1	Notation for 2-stage SAGP model.	27
4.2	Real-world aggregated data sets.	33
4.3	Performance comparisons for 2-stage SAGP model.	37
4.4	Ranking of auxiliary data sets.	40
5.1	Notation for SAGP-M.	47
5.2	Performance comparisons for SAGP-M.	54
5.3	Comparisons of computational time (second).	56
5.4	Performance comparisons for transfer learning across two cities.	58
6.1	Notation for T-CFDM.	68
6.2	Travel duration probabilities.	70
6.3	The totals of incoming and outgoing people in the real-world data sets. . . .	75
6.4	Mean normalized absolute errors (MNAEs) and standard errors of the estimates of transition populations.	78
6.5	Mean Kullback-Leibler divergence for the estimation of travel duration probability.	79

Chapter 1

Introduction

1.1 Motivation

1.1.1 Data and applications

Recent advances in information technology have increased the availability of data collected from various spaces such as urban cities, shopping malls, amusement parks, and exhibition halls. Organizations including governments and corporations are now gathering and releasing data on items such as poverty, air pollution, traffic, energy consumption, and crime (*NYC Open Data*; *Chicago Data Portal*). In general, each original data sample is represented by a pair of location points and values. The left part of Figure 1.1 shows an example of the data given five observed location points. In this thesis, this type of data is called *individual-level data*. We describe two instances of individual-level data as follows:

- **Poverty data.** Household surveys are conducted in most cities to examine socio-economic status. The surveys include several items such as income and the level of education; the poverty rate is calculated from these results. In this example, the value corresponds to the poverty rate represented by a positive number, which is associated with the household address.
- **Human location data.** One can determine an individual's location by wireless technologies including global positioning systems (GPS) and Bluetooth Low Energy (BLE). In this example, the value is assigned as 1 to indicate the existence of a person; each sample is obtained at the location of a sensor (e.g., mobile phone and BLE beacon).

Table 1.1 summarizes five examples of location points and values that are specified by the kind of individual-level data set.

These data play a crucial role in improving the life quality of people in many aspects including socio-economics (Rupasinghaa and Goetz, 2007; Smith, Mashhadi, and Capra, 2014), epidemiology (Sturrock et al., 2014), public security (Bogomolov et al., 2014; Wang et

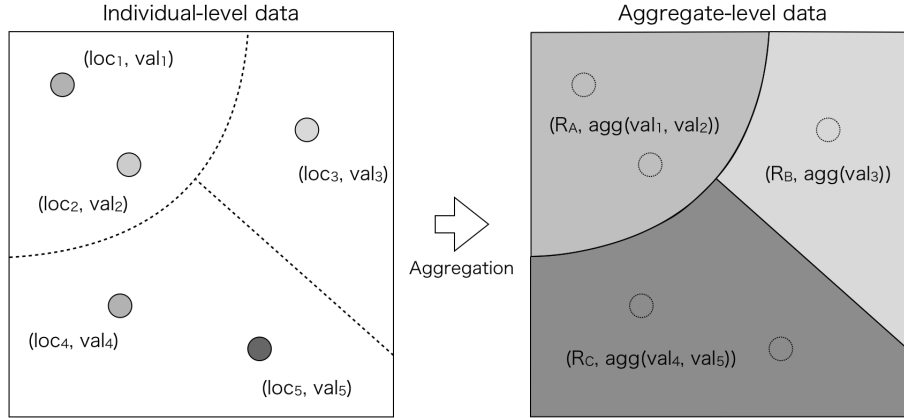


FIGURE 1.1: Data illustration. (Left figure) Illustration of individual-level data. Five gray dots represent observed location points, where darker hues represent points with higher values. Each sample is represented by a pair of location points and values. (Right figure) Illustration of aggregate-level data given three spatial regions R_A , R_B , R_C . $\text{agg}(\bullet)$ is the result of aggregation processes (e.g., spatial averaging), where the hue of each region is proportional to the resulting value. Each sample is associated with regions, not with location points.

TABLE 1.1: Examples of location point and value for each kind of individual-level data set. The value 1 indicates the existence of an observation.

Data	Location point	Value
Poverty data	Household address	Poverty rate
Air pollution data	Monitoring station location	Air quality index
Human location data	Location of sensor (e.g., mobile phone)	1
Crime data	Location where an incident happened	1
Disease data	Patient's address	1

al., 2016), public health (Jerrett et al., 2013), urban planning (Yuan, Zheng, and Xie, 2012), disaster management (Song et al., 2014), navigation systems (Huang and Gartner, 2010; Kurashima et al., 2010), and location-based marketing (Dhar and Varshney, 2011). For instance, knowing the spatial distribution of poverty in a city is important for identifying key regions that require intervention; this allows us to optimize the allocation of resources for remedial action. As another instance, capturing human mobility patterns in a city yields better route recommendations and helps determine policy for urban planning.

1.1.2 Challenges

In practice, individual-level data gathered from a space of interest are not only anonymized but also often *aggregated* over spatial regions. The resulting values are obtained via aggregation processes, e.g., averaging and summation, over the values associated with the

location points in each spatial region. The spatial regions are assumed to be predefined for each data set. For example, crime rate data may be associated with police precincts. Population data gathered by sensors (e.g., BLE beacons) may be identified at each sensor location, in which case the spatial regions correspond to the observation ranges of sensors. The right part of Figure 1.1 shows an example of the data given three spatial regions. In this thesis, this type of data is called *aggregate-level data* or simply *aggregated data*.

There are several practical reasons why the data should be aggregated. First, most of the data sets have, either explicitly or implicitly, personal information: Individual identifier, address, education, occupation, income, and so on. Thus, one should aggregate the data for preserving people’s privacy. Second, governments and other stakeholders want to know a status (e.g., a socio-economic status) at the district level, because the intervention is typically conducted in the district. For example, some type of resource allocation might be determined according to the level of poverty in each district rather than the economic conditions of the individuals living there.

Although data aggregation is an essential process for the above reasons, it makes it difficult to apply existing methods for spatial data analysis, most of which assume that individual-level data are available. The main goal of this thesis is to present a framework for modeling spatially aggregated data, which has the capability for prediction and knowledge discovery from aggregated data. For each specific problem to be discussed in this thesis, we propose a probabilistic model that incorporates a key mechanism, i.e., an aggregation process, for handling data aggregated over regions; this allows for effective model learning based on aggregated data. Our proposed model provides a computational lens for discovering knowledge needed by practitioners even if one has only aggregate-level data.

The specific problems addressed in the thesis are described in the next section.

1.2 Thesis overview and contributions

1.2.1 Overview

The thesis specifically addresses two important problems involved in handling aggregated data. Accordingly, the thesis has a two-part structure; it focuses on two problem domains: 1) refining coarse-grained aggregated data; 2) inferring latent people flow from aggregated population data. We propose a probabilistic model and its inference algorithm for solving each problem. Summary of each part is as follows.

1.2.2 Part 1 – Refining coarse-grained aggregated data

Understanding the spatial distribution of data, such as poverty rates and air pollution rates, in a city is needed to improve city environments efficiently. The data gathered from a city are actually associated with one of the predefined geographical partitions with various granularities including boroughs, community districts, zip code, and police precincts. We address the problem of refining coarse-grained aggregated data by utilizing multiple aggregated data sets with various granularities, which enables us to find pin-point key regions, e.g., the poorest regions of a city, efficiently. There have been no methods that can use multiple aggregated data sets with various granularities for refinement task.

We examine a formulation based on Gaussian processes (GPs) (Rasmussen and Williams, 2006) for refining coarse-grained aggregated data. The standard GP models assume that each sample is observed at a location point, not at a region. In order to handle the data aggregated over regions, we first formulate *Spatially Aggregated Gaussian Processes with a Single Output (SAGP-S)* in Chapter 3. In this formulation, we introduce an observation model with an aggregation process, which is an integral of a GP over the spatial region. This allows us to accurately evaluate variances and covariances with consideration of region shapes. Fine-grained data to be estimated can be obtained by deriving the posterior GP.

We propose two probabilistic models by extending SAGP-S to utilize multiple aggregated data sets with various granularities. The major difference between two proposed models is that dependences between data sets are established by different approaches: 1) regression approach and 2) multivariate approach. The main contributions of our models are summarized in the following approach-by-approach outline.

Regression approach

Originally published in the Proceedings of the 33rd AAAI Conference on Artificial Intelligence (Tanaka et al., 2019a).

In Chapter 4, we propose *Two-stage Spatially Aggregated Gaussian Process Model (2-stage SAGP model)* for refining coarse-grained aggregated data. In the regression approach, a target data to be refined is picked up; 2-stage SAGP model learns the regression coefficients for the auxiliary data sets. To handle auxiliary data sets with various granularities, we first derive a posterior GP separately for each auxiliary data set; this conceptually corresponds to spatial interpolation. The target data are assumed to be another GP that is modeled by a linear combination of the posterior GPs for auxiliary data sets. This allows us to learn the regression coefficients with consideration of the posterior variances for each

auxiliary data set. This is beneficial in that the usefulness of each auxiliary data set can be determined by not only the strength of relationship with the target data but also the uncertainty in spatial interpolation for each auxiliary data set. Since the target data are observed at regions, we introduce an observation model with an aggregation process, as in SAGP-S.

By analytically integrating out the GPs for target and auxiliary data sets, we can estimate the parameters on the basis of the marginal likelihood. Since the estimation procedure can be divided into two steps, it has an advantage in computational efficiency. By constructing the posterior GP for the target data, we can obtain the fine-grained target data.

Experiments on real-world data sets from urban cities demonstrate that 2-stage SAGP model can 1) accurately refine coarse-grained aggregated data than the existing regression-based models and 2) effectively assess the usefulness of each auxiliary data set considering the strength of relationships with target data and the posterior variances for the auxiliary data sets, simultaneously.

Multivariate approach

Originally published in the Proceedings of the 33rd Conference on Neural Information Processing Systems (Tanaka et al., 2019b).

In Chapter 5, we propose *Spatially Aggregated Gaussian Processes with Multiple Outputs (SAGP-M)*, a novel multivariate GP model that can infer the multivariate function from multiple aggregated data sets with various granularities. To establish the dependences between data sets, the multivariate approach aims to model a joint distribution of all data sets. In SAGP-M, the functions for respective data sets are assumed to be a multivariate dependent GP that is modeled as a linear mixing of independent latent GPs. Sharing of latent GPs across multiple data sets allows effective learning of spatial correlation for each data set; what is more, it can easily be used for transfer learning across multiple domains (e.g., cities). To handle the multiple aggregated data sets, we design an observation model with aggregation processes for all aggregated data sets, which is an integral of the mixed GP over the corresponding region.

We develop a parameter estimation procedure based on the marginal likelihood in which latent GPs are analytically integrated out. By deriving the posterior GP given multiple aggregated data sets, one can refine the coarse-grained aggregated data.

We conduct experiments on multiple real-world data sets from urban cities; the results show that SAGP-M can 1) accurately refine coarse-grained aggregated data and 2) improve refinement performance via the use of aggregated data sets from multiple cities.

1.2.3 Part 2 – Inferring latent people flow from aggregated population data

Originally published in the Proceedings of the 27th International Joint Conference on Artificial Intelligence (Tanaka et al., 2018).

Uncovering people’s movements in a space of interest is of critical importance in many applications such as navigation systems and urban planning. We assume the case where location information of people is aggregated over spatial regions (e.g., sensor ranges and grid cells) to preserve people’s privacy; then we have aggregated population data at each region. In Part 2, we simply call the region as *location*; it should be noted that it does not mean location points. In addition, we assume that the aggregated population data can be available at each time step. Since the aggregated data lack tracking information of individuals, determining the people flows between locations is not straightforward.

We address the problem of inferring latent people flows, that is, transition populations between locations, given just aggregated population data. The aggregated data we use as input consist of the numbers of incoming and outgoing people at each location and at each time step. Solving this problem enables us to analyze people’s mobility patterns while protecting privacy.

In practical situations, it is unrealistic to always observe everyone in at least one of the observed locations. Consider the case where location data are gathered by sensors (e.g., BLE beacons). We do not always have a large enough number of sensor devices to cover a large-scale space targeted; the observation range is limited and some people are not observed in any location in some time periods. In such situations, existing models fail to estimate transition populations between locations accurately.

In Chapter 6, we propose a probabilistic model, *Time-delayed Collective Flow Diffusion Models (T-CFDM)*, that robustly infers transition populations between locations from aggregated population data gathered at limited locations. In T-CFDM, we assume that individuals move between locations in accordance with time-independent transition probabilities that are shared among people. Since the transition populations between locations are not obtained, we treat them as latent variables. We introduce an observation model with an aggregation process to encourage people flow conservation. A key idea is to incorporate travel duration distributions between observed locations into constraints that represent flow conservation law, which yields an accurate estimation of transition populations even if we have a limited number of observed locations.

We develop an efficient approximate expectation-maximization algorithm that can estimate transition populations and model parameters, simultaneously.

By using multiple real-world data sets: pedestrian location logs from large-scale exhibition halls, and bike trip data and taxi trip data from an urban city, we confirm that T-CFDM achieves better estimation performance than the compared methods.

1.3 Mathematical notations

In this section, we describe mathematical notations that we will use throughout this thesis. Vectors are denoted by lower-case bold letters, e.g., $\mathbf{x}$, where all vectors are assumed to be column vectors. $\mathbf{x}^\top$ denote the transpose of vector $\mathbf{x}$. Matrices or sets are denoted by capital bold letters, e.g., $\mathbf{X}$, and $|\mathbf{X}|$ denotes the number of elements in set $\mathbf{X}$.

$\mathbf{I}$ denotes the identity matrix of appropriate dimension. The function $\mathbb{1}(\bullet)$ denotes the indicator function represented by

$$\mathbb{1}(B) = \begin{cases} 1 & \text{if } B \text{ is true,} \\ 0 & \text{otherwise.} \end{cases} \quad (1.1)$$

$\delta_{\bullet,\bullet}$ denotes Kronecker's delta represented by

$$\delta_{B,B'} = \begin{cases} 1 & \text{if } B = B' \\ 0 & \text{otherwise.} \end{cases} \quad (1.2)$$

$\mathbb{E}[\bullet]$ denotes expectation.

Part I

Refining Coarse-grained Aggregated Data

Chapter 2

Introduction and survey

2.1 Introduction

In many cities around the world, a wide variety of aggregated data sets are recently available (Shadbolt et al., 2012; Goldstein and Dyson, 2013; Barlacchi et al., 2015). These data sets are essential for improving city environments in many disciplines such as socio-economics (Rupasinghaa and Goetz, 2007; Smith, Mashhadi, and Capra, 2014), public security (Bogomolov et al., 2014; Wang et al., 2016), public health (Jerrett et al., 2013), and urban planning (Yuan, Zheng, and Xie, 2012). For example, the poverty distribution in a city is helpful for optimizing resource allocation for remedial action. Likewise, we can use the air pollution distribution for helping create policies that can control air quality.

In order to achieve the effective improvements of city environments, it is important to obtain the data at finer granularity; the fine-grained data allows us to easily identify pinpoint key regions that require intervention. Figures 2.1(a) and 2.1(b) illustrate the visualization examples of poverty rate data in New York City by community district and by borough, respectively, where darker hues represent poorer regions. Figure 2.1(a) is preferred than Figure 2.1(b) to better understand socio-economic problems in the city. In practice, however, such information is often aggregated into coarser granularities, as in Figure 2.1(b). It is indeed difficult to collect the large amount of individual-level data via a *household survey* over the whole population of a city, for the survey is too time-consuming and costly; one often opts for a *sample survey* instead. That is why the number of samples associated with each fine-grained region may not be large enough to provide a statistically significant estimate of the value associated to this region. One has no choice but to aggregate samples over larger regions (Smith, Mashhadi, and Capra, 2014).

In this part, we tackle the problem of refining coarse-grained aggregated data. One promising approach to address this problem is to utilize a wide variety of aggregated data sets

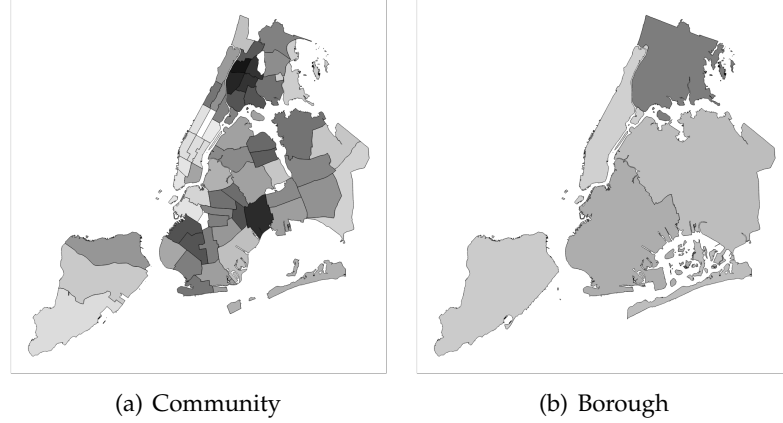


FIGURE 2.1: The spatial distribution of poverty in New York City at different spatial granularities.

from the same domain (e.g., a city). Existing models have been based on at least one of the following assumptions:

- All aggregated data sets, except for a coarse-grained target data set, have sufficiently fine spatial granularity (e.g., $1 \text{ km} \times 1 \text{ km}$ grid cells).
- The data samples are observed at location points, not at regions; namely, we have the individual-level data.

Unfortunately, many aggregated data sets are actually associated with relatively large regions (e.g., zip code and police precinct); these existing models are not straightforwardly applicable in such situations.

We propose two probabilistic models based on Gaussian processes (GPs) (Rasmussen and Williams, 2006) for refining coarse-grained aggregated data by using multiple aggregated data sets with various granularities. Both proposed models incorporate a mechanism, a spatial aggregation process, essential for handling data that have been aggregated over regions, which is represented by an integral of a GP over the corresponding region. The major difference between the two proposed models is that dependences between data sets are established by different approaches: 1) regression approach and 2) multivariate approach, each of which is described in Chapters 4 and 5.

The remainder of this chapter is organized as follows. Section 2.2 provides the related works that are categorized into either 1) regression approach or 2) multivariate approach. Section 2.3 describes mathematical notations for spatially aggregated data used in this part and defines our problem of refining coarse-grained aggregated data. Finally, we present a roadmap of this part in Section 2.4.

2.2 Related work

2.2.1 Categorization

Related works can be roughly categorized into two approaches: 1) regression approach and 2) multivariate approach. The major difference between them is as follows: Denoting $\mathbf{y}^t$ and $\mathbf{y}^a$ as target data and auxiliary data, respectively, the aim of the first approach is to design a conditional distribution $p(\mathbf{y}^t \mid \mathbf{y}^a)$; the second approach designs a joint distribution $p(\mathbf{y}^t, \mathbf{y}^a)$ instead.

2.2.2 Regression approach

A related problem has been addressed in the geostatistics community under the name of *downscaling*, *disaggregation*, *areal interpolation*, or *fine-scale modeling* (Gotway and Young, 2002), and this has attracted great interest in many disciplines such as socio-economics (Bogomolov et al., 2014; Smith, Mashhadi, and Capra, 2014), agricultural economics (Howitt and Reynaud, 2003; Xavier et al., 2016), epidemiology (Sturrock et al., 2014), meteorology (Wotling et al., 2000; Wilby et al., 2004), and geographical information systems (Goovaerts, 2010). The regression-based models can further be categorized into two cases in terms of target data availability.

The first case: A large amount of coarse- and fine-grained target data are available for training. In that case, one can predict the fine-grained target data by training a mapping function from coarse- to fine-grained data, in which mapping function is learned by using various statistical methods including linear regression models (Hessami et al., 2008), neural networks (Cannon, 2011; Misra, Sarkar, and Mitra, 2017), and support vector machines (Ghosh, 2010). Recently, super-resolution techniques based on deep neural networks have been applied for refining coarse-grained aggregated data (Vandal et al., 2017; Vandal et al., 2018). The super-resolution techniques aim to learn a mapping function from low- to high-resolution images (Dong et al., 2014). The method by (Vandal et al., 2017) is based on the analogy between data aggregated over grid cells and images; values at grid cells are regarded as values at pixels. However, the large amount of fine-grained aggregated data needed for training are not available in many practical cases (e.g., poverty survey), and often only coarse-grained aggregated data are available. These methods are not applicable in such situations.

The second case: Only coarse-grained target data are available, which is the focus of this study. Regression models (linear and non-linear) are used for learning the regression coefficients of target data for auxiliary data sets that have fine granularity (e.g., 1 km

$\times 1$ km grid cells) (Diodato et al., 2010; Zheng, Liu, and Hsieh, 2013; Smith, Mashhadi, and Capra, 2014; Zheng et al., 2015; Flaxman, Wang, and Smola, 2015; Miller et al., 2015; Wang et al., 2016; Xu, 2017; Xu et al., 2018). In the models of (Murakami and Tsutsumi, 2011; Park, 2013), the regression coefficients can be estimated under the spatial aggregation constraints that encourage consistency between fine- and coarse-grained target data. The aggregation constraints have been incorporated via block kriging (Burgess and Webster, 1980) or transformations of Gaussian process (GP) priors (Murray-Smith and Pearlmutter, 2004; Smith, Álvarez, and Lawrence, 2018), which are developed for addressing the *change of support problem* in geostatistics. There have been a number of advanced models that offer a fully Bayesian inference (Keil et al., 2013; Taylor, Andrade-Pacheco, and Sturrock, 2018; Wilson and Wakefield, 2018) or a variational inference (Law et al., 2018) for model parameters.

These models require that all auxiliary data sets are either the individual-level data or the data aggregated over sufficiently fine-grained regions, for one needs to prepare the auxiliary data sets that have spatial granularities equivalent to that of fine-grained target data to be estimated. However, the aggregated data available on a city are actually associated with various geographical partitions, thus one might not be able to obtain the auxiliary data sets that meet the above requirements. Accordingly, these models are not applicable for multiple auxiliary data sets with various granularities.

In Chapter 4, we propose a new regression-based model that can use auxiliary data sets with various granularities to refine coarse-grained target data. Our model is based on the following three parts: 1) deriving the posterior GPs for auxiliary data sets separately; 2) modeling a GP for target data as a linear combination of the posterior GPs for auxiliary data sets; 3) introducing an observation model with spatial aggregation processes for target data. This formulation allows for supporting multiple auxiliary data sets with various granularities and learning model parameters while encouraging the aggregation constraints as in block kriging.

2.2.3 Multivariate approach

Multivariate modeling is another option for establishing the dependences between data sets, which is generally advantageous as it can alleviate the data sparsity problem. In geostatistics, the classical method of *co-kriging* is widely used for predicting multivariate spatial data (Myers, 1984); this method is, however, problematic in that it is unclear how to define cross-covariance functions that determine the dependences between data sets (Gibbs and MacKay, 1997). The linear model of coregionalization (LMC) is one of the

most widely-used approaches for constructing a multivariate function; the dependent outputs are expressed as a linear mixing of independent latent functions (Álvarez, Rosasco, and Lawrence, 2012). In the machine learning literature, there has been growing interest in multivariate GPs (Rasmussen and Williams, 2006), in which dependences between data sets are introduced via methodologies such as process convolution (Higdon, 2002; Boyle and Frean, 2005), latent factor modeling (Teh, Seeger, and Jordan, 2005; Luttinen and Ilin, 2009), and multi-task learning (Micchelli and Pontil, 2004; Bonilla, Chai, and Williams, 2008). The semiparametric latent factor model (SLFM) is an instance of LMC, in which latent functions are defined by GPs (Teh, Seeger, and Jordan, 2005). Unfortunately, these multivariate GP models cannot be straightforwardly used for modeling the aggregated data we focus on, because they assume that the data samples are observed at location points; namely they do not have a mechanism, a spatial aggregation process, essential for handling data that have been aggregated over regions.

In Chapter 5, we propose a novel probabilistic model for inferring multivariate function from multiple aggregated data sets with various granularities. Our model is an extension of SLFM. To handle the multiple aggregated data sets, we introduce an observation model with the spatial aggregation processes for all data sets, which is represented by the integral of the mixed GP over each corresponding region. The sharing of key information (i.e., covariance function) can be used for transfer learning across a wide variety of data sets, which allows our model 1) to robustly estimate the spatial correlations for the respective data sets and 2) to utilize aggregated data sets from multiple domains (e.g., cities).

Multi-task GP models have recently and independently been proposed for addressing similar problems (Hamelijnck et al., 2019; Yousefi, Smith, and Álvarez, 2019). Main differences of our work from them are as follows: 1) explicit derivation of the posterior GP given multivariate aggregated data; 2) transfer learning across multiple domains; 3) extensive experiments on real-world data sets defined on the two-dimensional input space.

2.3 Problem formulation

In this section, we describe the mathematical notations used in this part, and define the problem of refining coarse-grained aggregated data. Assume that we have a wide variety of aggregated data sets from the same domain (e.g., a city), and each data set is associated with one of the geographical partitions that have various granularities. Let S be the number of kinds of data sets. The notations are listed in Table 2.1.

Partition and region. Let $X \subset \mathbb{R}^2$ denote an input space (e.g., a total region of a city), and $\mathbf{x} \in X$ denote an input variable, represented by its coordinates (e.g., latitude and

TABLE 2.1: Notation for spatially aggregated data.

Symbol	Description
S	number of kinds of data sets
s	data set index, $s = 1, \dots, S$
X	input space, $X \subset \mathbb{R}^2$
x	input variable, $x \in X$
P_s	partition of X for the s -th data set
$ P_s $	number of regions in P_s
n	region index, $n = 1, \dots, P_s $
$R_{s,n}$	the n -th region in P_s
$y_{s,n}$	value associated with $R_{s,n}$, $y_{s,n} \in \mathbb{R}$
D_s	the s -th aggregated data set, $D_s = \{(R_{s,n}, y_{s,n}) \mid n = 1, \dots, P_s \}$
D	S aggregated data sets, $D = \cup_{s=1}^S D_s$
P^{fine}	target fine-grained partition of X
R_n^{fine}	the n -th region in P^{fine}
$y_{s,n}^{\text{fine}}$	target fine-grained data value associated with R_n^{fine} , $y_{s,n}^{\text{fine}} \in \mathbb{R}$

longitude). For $s = 1, \dots, S$, the partition P_s of X is a collection of disjoint subsets, called *regions*, of X . Let $|P_s|$ be the number of regions in P_s . For $n = 1, \dots, |P_s|$, let $R_{s,n} \in P_s$ denote the n -th region in P_s .

Spatially aggregated data. Let $y_{s,n} \in \mathbb{R}$ denote a value associated with the n -th region $R_{s,n}$ of the s -th data set. Each data sample is represented by the pair $(R_{s,n}, y_{s,n})$. Suppose that we have S aggregated data sets $D = \cup_{s=1}^S D_s$, where $D_s = \{(R_{s,n}, y_{s,n}) \mid n = 1, \dots, |P_s|\}$.

Problem. Let P^{fine} be the target fine-grained partition, and $R_n^{\text{fine}} \in P^{\text{fine}}$ be the n -th region in P^{fine} . Suppose that we have the aggregated data sets D and the target partition P^{fine} , we would like to predict the fine-grained data $\{y_{s,n}^{\text{fine}} \in \mathbb{R} \mid n = 1, \dots, |P^{\text{fine}}|\}$ associated with the target regions $\{R_n^{\text{fine}} \mid n = 1, \dots, |P^{\text{fine}}|\}$ for an arbitrary data set index s .

2.4 Roadmap

The remainder of this part is organized as follows. In Chapter 3, we first describe *Spatially Aggregated Gaussian Processes with a Single Output (SAGP-S)*, which is the GP formulation we developed for handling the data aggregated over regions. This is a key technique for constructing the following proposed models.

In Chapter 4, we propose a regression-based model, *Two-stage Spatially Aggregated Gaussian Process Model (2-stage SAGP model)*, for refining coarse-grained aggregated data by using

multiple auxiliary data sets with various granularities. This model is based on three parts: 1) the posterior of SAGP-S is separately obtained for each auxiliary data set; 2) a GP for target data is modeled by a linear combination of the posterior GPs for auxiliary data sets; 3) introducing an observation model with the output represented by an integral of the GP for target data. 2-stage SAGP model adopts a two-step inference procedure for learning model parameters, which has an advantage in computational efficiency.

In Chapter 5, we propose a multivariate GP model, *Spatially Aggregated Gaussian Processes with Multiple Outputs (SAGP-M)*, which is a natural extension of SAGP-S to handle multiple outputs. Multivariate dependent GP is constructed by a linear mixing of independent latent GPs. We design an observation model with spatial aggregation processes for the multivariate aggregated data. A one-step inference procedure for estimating model parameters is available in SAGP-M; this allows to effectively estimate spatial correlations for respective data sets by alleviating the data sparsity problems involved in coarse-grained aggregated data.

Chapter 3

Spatially Aggregated Gaussian Processes with a Single Output

3.1 Overview

In this chapter, we describe a probabilistic model, *Spatially Aggregated Gaussian Processes with a Single Output (SAGP-S)*, for inferring a function from a single aggregated data set, which is a core idea for constructing the models proposed in Chapters 4 and 5. Gaussian process (GP) is widely used as a prior over the continuous function on the input space (Rasmussen and Williams, 2006). Most of GP-based models assume that the samples are identified by location points, not by regions; thus, these GP-based models cannot straightforwardly apply to modeling of data aggregated over regions. To address this issue, block kriging (Burgess and Webster, 1980) and transformations of GP priors (Murray-Smith and Pearlmutter, 2004; Smith, Álvarez, and Lawrence, 2018) have been developed. A common idea of these works is to model the aggregated output as an integral of the random function (e.g., GP) over the appropriate integral interval.

In the remainder of this chapter, we present the formulation of SAGP-S and its inference procedures. Although our formulation is based on the idea of prior works (Burgess and Webster, 1980; Murray-Smith and Pearlmutter, 2004; Smith, Álvarez, and Lawrence, 2018), it is more informative in that 1) we discuss the integrability of GP and 2) we describe the explicit derivation of marginal likelihood and posterior GP given aggregated data, which allows the formulation of our proposed models in Chapters 4 and 5.

3.2 Model

Consider the case of a single output; namely we obtain just one aggregated data set. Let $\{(\mathbf{R}_n, y_n) \mid n = 1, \dots, |P|\}$ denote the aggregated data, each value $y_n \in \mathbb{R}$ of which is

associated with the n -th region R_n in the partition P .

In a way similar to the standard GP approach, a noise-free latent function $f(x) : X \rightarrow \mathbb{R}$ for observed data is assumed to be a GP represented by

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \quad (3.1)$$

where $m(x) : X \rightarrow \mathbb{R}$ and $k(x, x') : X \times X \rightarrow \mathbb{R}$ are a mean function and a covariance function, respectively, both of which are assumed integrable. The mean function $m(x)$ defines the mean value of function $f(x)$ at location point x . The covariance function $k(x, x')$ determines the spatial correlation between the pair of location points (x, x') .

To handle aggregated observations, we introduce an observation model with a spatial aggregation process. Let $\mathbf{y} = (y_1, \dots, y_{|P|})^\top$ be a $|P|$ -dimensional vector consisting of the aggregated observations. Each observation is assumed to be obtained by integrating the GP $f(x)$ over the corresponding region; $\mathbf{y}$ is generated from the Gaussian distribution¹,

$$\mathbf{y} | f(x) \sim \mathcal{N}\left(\mathbf{y} \mid \int_X \mathbf{a}(x) f(x) dx, \sigma^2 \mathbf{I}\right), \quad (3.2)$$

where $\mathbf{a}(x) = (a_1(x), \dots, a_{|P|}(x))^\top$, whose entry $a_n(x)$ is a nonnegative weight function for spatial aggregation over region R_n . This formulation does not depend on the particular choice of $\{a_n(x)\}$, provided that they are integrable. If one takes, for region R_n ,

$$a_n(x) = \frac{\mathbb{1}(x \in R_n)}{\int_X \mathbb{1}(x' \in R_n) dx'}, \quad (3.3)$$

then y_n is the average of $f(x)$ over R_n . One may also consider other aggregation processes to suit the property of the aggregated observations, including simple summation and population-weighted averaging over R_n . Here, σ^2 in (3.2) is the noise variance for the observation model.

In our formulation, the observations are assumed to be obtained from the Gaussian distribution (3.2). This option has the advantage that marginalization of GPs for inference can be analytically calculated; the details are contained in the next section. Another option is to opt for Poisson distribution for count data. In that case, we should employ approximate inference algorithms, e.g., variational inference, as in (Law et al., 2018).

¹We here assume that the integral appearing in (3.2) is well-defined. Actually, sample paths of a Gaussian process are in general not integrable without additional assumptions; the conditions under which the integral is well-defined are discussed in Appendix A.

3.3 Inference

3.3.1 Outline

Given the aggregated data $\mathbf{y}$, we attempt to derive the posterior GP defined on $\mathbf{X}$ on the basis of a Bayesian inference procedure. The model parameters, $m(\mathbf{x})$, $k(\mathbf{x}, \mathbf{x}')$, σ^2 , are learned by maximizing the marginal likelihood, in which GP $f(\mathbf{x})$ can be analytically integrated out; we then construct the posterior GP by using the estimated parameters. Once the posterior GP is obtained, one can predict a data value at any location point. The predicted value $y_{s,n}^{\text{fine}}$ associated with the target fine-grained region $\mathbf{R}_n^{\text{fine}}$ can be calculated by integrating the posterior GP over the region $\mathbf{R}_n^{\text{fine}}$.

3.3.2 Marginal likelihood

Given the aggregated data $\mathbf{y}$, the marginal likelihood is given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{C}), \quad (3.4)$$

where we analytically integrate out the GP prior $f(\mathbf{x})$. Here, $\boldsymbol{\mu}$ is a $|\mathbf{P}|$ -dimensional mean vector represented by

$$\boldsymbol{\mu} = \int_{\mathbf{X}} \mathbf{a}(\mathbf{x}) m(\mathbf{x}) d\mathbf{x}, \quad (3.5)$$

which is the integral of the mean function $m(\mathbf{x})$ over the respective regions. $\mathbf{C}$ is a $|\mathbf{P}| \times |\mathbf{P}|$ covariance matrix represented by

$$\mathbf{C} = \iint_{\mathbf{X} \times \mathbf{X}} k(\mathbf{x}, \mathbf{x}') \mathbf{a}(\mathbf{x}) \mathbf{a}(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \sigma^2 \mathbf{I}. \quad (3.6)$$

The derivation of the marginal likelihood (3.4) is described in the following paragraph. Equation (3.6) provides the region-to-region covariance matrix in the form of the double integral of the covariance function $k(\mathbf{x}, \mathbf{x}')$ over the respective pairs of regions in $\mathbf{P} \times \mathbf{P}$; this conceptually corresponds to aggregation of the covariance function values that are calculated at the infinite pairs of location points contained in the corresponding region pairs. Since the integrals over regions in (3.5) and (3.6) cannot be calculated analytically, in practice we use a numerical approximation of these integrals. Details are provided in Section 3.3.4. The formulation (3.6) is advantageous in the following reasons.

- The variances at regions can be appropriately evaluated considering the region sizes. In the case of region-average model (3.3), the larger the region size is, the smaller the

variance is. This is beneficial functionality because there might be the regions with various sizes in the input aggregated data.

- This allows for accurately evaluating the covariances between regions considering their shapes; this is extremely helpful as some input data are likely to originate from irregularly shaped regions (e.g., extremely elongated).

By maximizing the logarithm of the marginal likelihood (3.4), we can estimate the model parameters. The optimization problem can be solved by using any numerical optimization method such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Liu and Nocedal, 1989).

Derivation of the marginal likelihood (3.4)

The integral of the GP $f(\mathbf{x})$ appearing in (3.2) is assumed to be well-defined (See Appendix A for the integrability condition). Let $\mathbf{x}_1, \dots, \mathbf{x}_I$ denote I grid points of a regular grid covering $\mathbf{X}$, and Δ denote the grid cell volume. We consider Riemann sums to approximate the integral in (3.2) on $\mathbf{X}$; and analytically integrate out the GP prior $f(\mathbf{x})$. We then take the limit $\Delta \rightarrow 0$ to derive the marginal likelihood (3.4).

Let $\bar{\mathbf{f}} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_I))^\top$ be an I -dimensional vector consisting of the GP outputs at I grid points. By the definition of GP, $\bar{\mathbf{f}}$ is the I -dimensional Gaussian random variable; the distribution of $\bar{\mathbf{f}}$ is given by

$$\bar{\mathbf{f}} \sim \mathcal{N}(\bar{\mathbf{f}} \mid \bar{\mathbf{m}}, \bar{\mathbf{K}}), \quad (3.7)$$

where $\bar{\mathbf{m}} = (m(\mathbf{x}_1), \dots, m(\mathbf{x}_I))^\top$, and where $\bar{\mathbf{K}}$ is an $I \times I$ covariance matrix, whose (i, i') -entry is $k(\mathbf{x}_i, \mathbf{x}_{i'})$. Consider the observation process yielding the aggregated observations $\mathbf{y}$ defined by

$$\mathbf{y} \mid \bar{\mathbf{f}} \sim \mathcal{N}(\mathbf{y} \mid \bar{\mathbf{a}}\bar{\mathbf{f}}, \sigma^2 \mathbf{I}), \quad (3.8)$$

where $\bar{\mathbf{a}} = (\mathbf{a}(\mathbf{x}_1), \dots, \mathbf{a}(\mathbf{x}_I)) \Delta$. Equation (3.8) shows that $\mathbf{y}$ is a linear transformation of the Gaussian random variable; thus, the marginal likelihood of $\mathbf{y}$ approximated by Riemann sums is given by

$$\begin{aligned} p(\mathbf{y}) &= \int \mathcal{N}(\mathbf{y} \mid \bar{\mathbf{a}}\bar{\mathbf{f}}, \sigma^2 \mathbf{I}) \mathcal{N}(\bar{\mathbf{f}} \mid \bar{\mathbf{m}}, \bar{\mathbf{K}}) d\bar{\mathbf{f}} \\ &= \mathcal{N}(\mathbf{y} \mid \bar{\boldsymbol{\mu}}, \bar{\mathbf{C}}). \end{aligned} \quad (3.9)$$

Here, the mean vector $\bar{\boldsymbol{\mu}}$ and the covariance matrix $\bar{\mathbf{C}}$ are given by

$$\bar{\boldsymbol{\mu}} = \bar{\mathbf{a}}\bar{\mathbf{m}}, \quad (3.10)$$

$$\bar{\mathbf{C}} = \bar{\mathbf{a}}\bar{\mathbf{K}}\bar{\mathbf{a}}^\top + \sigma^2\mathbf{I}, \quad (3.11)$$

respectively.

By regarding sums over the I terms as Riemann sums approximating the corresponding integrals over $\mathbf{X}$, in the limit $\Delta \rightarrow 0$, one can replace those sums over I terms with the corresponding integrals over $\mathbf{X}$. The observation model of $\mathbf{y}$ is obtained as follows:

$$\begin{aligned} p(\mathbf{y}) &= \mathcal{N}\left(\mathbf{y} \mid \sum_{i=1}^I \mathbf{a}(\mathbf{x}_i)f(\mathbf{x}_i)\Delta, \sigma^2\mathbf{I}\right) \\ &\rightarrow \mathcal{N}\left(\mathbf{y} \mid \int_{\mathbf{X}} \mathbf{a}(\mathbf{x})f(\mathbf{x})d\mathbf{x}, \sigma^2\mathbf{I}\right). \end{aligned} \quad (3.12)$$

The mean vector and the covariance matrix are obtained as follows:

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= \sum_{i=1}^I \mathbf{a}(\mathbf{x}_i)m(\mathbf{x}_i)\Delta \\ &\rightarrow \int_{\mathbf{X}} \mathbf{a}(\mathbf{x})m(\mathbf{x})d\mathbf{x} = \boldsymbol{\mu}, \end{aligned} \quad (3.13)$$

$$\begin{aligned} \bar{\mathbf{C}} &= \sum_{i,i'=1}^I k(\mathbf{x}_i, \mathbf{x}_{i'})\mathbf{a}(\mathbf{x}_i)\mathbf{a}(\mathbf{x}_{i'})^\top \Delta^2 + \sigma^2\mathbf{I} \\ &\rightarrow \iint_{\mathbf{X} \times \mathbf{X}} k(\mathbf{x}, \mathbf{x}')\mathbf{a}(\mathbf{x})\mathbf{a}(\mathbf{x}')^\top d\mathbf{x}d\mathbf{x}' + \sigma^2\mathbf{I} = \mathbf{C}, \end{aligned} \quad (3.14)$$

showing that the mean vector $\bar{\boldsymbol{\mu}}$ and the covariance matrix $\bar{\mathbf{C}}$ of $\mathbf{y}$ are reduced in this limit to the vector $\boldsymbol{\mu}$ and the matrix $\mathbf{C}$ defined in (3.5) and (3.6), respectively.

3.3.3 Posterior GP

Given the aggregated data $\mathbf{y}$ and the estimated parameters, the posterior GP $f^*(\mathbf{x})$ is given by

$$f^*(\mathbf{x}) \sim \mathcal{GP}\left(m^*(\mathbf{x}), k^*(\mathbf{x}, \mathbf{x}')\right), \quad (3.15)$$

where $m^*(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}$ and $k^*(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ are the mean function and the covariance function for $f^*(\mathbf{x})$, respectively. Defining $\mathbf{h}(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}^{|P|}$ as

$$\mathbf{h}(\mathbf{x}) = \int_{\mathbf{X}} \mathbf{a}(\mathbf{x}')k(\mathbf{x}', \mathbf{x})d\mathbf{x}', \quad (3.16)$$

which consists of the point-to-region covariances; that is, the covariances between any location point $\mathbf{x}$ and the respective regions, the mean function $m^*(\mathbf{x})$ and the covariance function $k^*(\mathbf{x}, \mathbf{x}')$ are given by

$$m^*(\mathbf{x}) = m(\mathbf{x}) + \mathbf{h}(\mathbf{x})^\top \mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \quad (3.17)$$

$$k^*(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{h}(\mathbf{x})^\top \mathbf{C}^{-1} \mathbf{h}(\mathbf{x}'), \quad (3.18)$$

respectively. Derivation of the posterior GP (3.15) is detailed in the following paragraph. One can predict the data value at any location point by using the mean function (3.17); the data value $y_{s,n}^{\text{fine}}$ at the fine-grained region can be obtained by integrating the mean function (3.17) over the corresponding region $\mathbf{R}_n^{\text{fine}}$. By using the covariance function (3.18), one can also evaluate the prediction uncertainty.

Derivation of the posterior GP (3.15)

Letting $\mathbf{x}$ denote an arbitrary input variable, a joint distribution of $f(\mathbf{x})$ and $\mathbf{y}$ is given by a $(|\mathbf{P}| + 1)$ -dimensional Gaussian distribution,

$$\begin{bmatrix} \mathbf{y} \\ f(\mathbf{x}) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \bar{\boldsymbol{\mu}} \\ m(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} \bar{\mathbf{C}} & \bar{\mathbf{h}}(\mathbf{x})^\top \\ \bar{\mathbf{h}}(\mathbf{x}) & k(\mathbf{x}, \mathbf{x}) \end{bmatrix} \right), \quad (3.19)$$

where we use the marginal likelihood (3.8) approximated by Riemann sums, and where

$$\bar{\mathbf{h}}(\mathbf{x}) = \bar{\mathbf{a}} \bar{\mathbf{k}}. \quad (3.20)$$

Here, $\bar{\mathbf{k}}$ is an I -dimensional vector whose i -th element is $k(\mathbf{x}_i, \mathbf{x})$. The posterior of $f(\mathbf{x})$ given $\mathbf{y}$ is known to be a Gaussian distribution with mean

$$\bar{m}^*(\mathbf{x}) = m(\mathbf{x}) + \bar{\mathbf{h}}(\mathbf{x})^\top \bar{\mathbf{C}}^{-1}(\mathbf{y} - \bar{\boldsymbol{\mu}}) \quad (3.21)$$

and covariance

$$\bar{k}^*(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \bar{\mathbf{h}}(\mathbf{x})^\top \bar{\mathbf{C}}^{-1} \bar{\mathbf{h}}(\mathbf{x}'), \quad (3.22)$$

respectively. In the limit $\Delta \rightarrow 0$, one has

$$\begin{aligned} \bar{\mathbf{h}}(\mathbf{x}) &= \sum_{i=1}^I \mathbf{a}(\mathbf{x}_i) k(\mathbf{x}_i, \mathbf{x}) \Delta \\ &\rightarrow \int_X \mathbf{a}(\mathbf{x}') k(\mathbf{x}', \mathbf{x}) d\mathbf{x}' = \mathbf{h}(\mathbf{x}), \end{aligned} \quad (3.23)$$

where $h(\mathbf{x})$ is defined in (3.16). The above calculation shows that in the limit $\Delta \rightarrow 0$ the posterior process $f^*(\mathbf{x})$ is a GP with mean function $m^*(\mathbf{x})$ and covariance function $k^*(\mathbf{x}, \mathbf{x}')$ given by (3.17) and (3.18), respectively.

3.3.4 Approximation of the integral over regions

The integrals over regions in (3.5), (3.6), and (3.16) cannot be performed analytically; thus we approximate these integrals by using sufficiently fine-grained square grid cells. We divide input space $\mathbf{X}$ into square grid cells, and take $\mathbf{G}_n$ to be the set of grid points that are contained in region $\mathbf{R}_n$. Let us consider the approximation of the integral in the covariance matrix (3.6). The (n, n') -entry $C(n, n')$ of $\mathbf{C}$ is approximated as follows:

$$C(n, n') = \iint_{\mathbf{X} \times \mathbf{X}} k(\mathbf{x}, \mathbf{x}') a_n(\mathbf{x}) a_{n'}(\mathbf{x}') d\mathbf{x} d\mathbf{x}' + \delta_{n, n'} \sigma^2 \quad (3.24)$$

$$\approx \frac{1}{|\mathbf{G}_n|} \frac{1}{|\mathbf{G}_{n'}|} \sum_{i \in \mathbf{G}_n} \sum_{i' \in \mathbf{G}_{n'}} k(\mathbf{x}_i, \mathbf{x}_{i'}) + \delta_{n, n'} \sigma^2, \quad (3.25)$$

where we use the formulation of the region-average model (3.3). The integrals in (3.5) and (3.16) can be approximated in a similar way. Letting $|\mathbf{G}|$ denote the number of all grid points, the computational complexity of $\mathbf{C}$ (3.6) is $O(|\mathbf{G}|^2)$; assuming the constant weight $a_n(\mathbf{x}) = a_n$ (e.g., region average), the computational complexity can be reduced to $O(|\mathbf{P}||\mathbf{P}||\mathbf{Q}|)$, where $|\mathbf{Q}|$ is the cardinality of the set of distinct distance values between grid points. Here, we use the property that $k(\mathbf{x}_i, \mathbf{x}_{i'})$ in (3.25) depends only on the distance between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$. This is useful for reducing the computation time and the memory requirement.

Chapter 4

Regression approach

4.1 Overview

In this chapter, we discuss the methods based on the regression approach. As described in Section 2.2.2, the existing regression-based models are applicable under the assumption that we obtain the auxiliary data sets that have the spatial granularity equivalent to that of the target fine-grained partition; thus, these models cannot utilize multiple auxiliary data sets with various granularities.

We propose a probabilistic model, *Two-stage Spatially Aggregated Gaussian Process Model (2-stage SAGP model)*, for refining coarse-grained target data via the use of multiple auxiliary data sets with various granularities. In 2-stage SAGP model, we adopt the regression approach for estimating the relationships between target data and auxiliary data sets. Since the auxiliary data sets have various granularities, we first apply SAGP-S (described in Chapter 3) to each auxiliary data set separately to derive the posterior Gaussian process (GP) defined on the continuous space; this conceptually corresponds to spatial interpolation. A key idea is that the target data are assumed to be another GP that is modeled by a linear combination of the posterior GPs for auxiliary data sets, in which posterior GPs are analytically integrated out; it should be noted that it does not use point estimates. This allows us to consider the posterior variances for each auxiliary data set to determine their usefulness. The variances represent the uncertainty in spatial interpolation, which is governed by several factors; for example, they might be affected by sample density, that is, spatial granularity of auxiliary data; the finer the granularity is, the lower the variance is. As another example, the variances might be affected by the degree of stationarity of data; if the auxiliary data have actually a highly non-stationary property and the covariance function is assumed to be stationary, then the estimated variances tend to be higher. Accordingly, these variances play an important role in learning the usefulness of each auxiliary data set that have various granularities. Since the target data are observed not at

location points but at regions, we also adopt an observation model for target data with a spatial aggregation process, as in SAGP-S.

We obtain the posterior GP for target data via a Bayesian inference procedure, where the parameter estimation can be performed by maximizing the marginal likelihood in which GP for target data are analytically integrated out. The estimation of 2-stage SAGP model is computationally efficient because its procedure can be divided into two steps.

By conducting experiments on multiple real-world data sets from New York City and Chicago, we confirm that our model can refine the coarse-grained target data more accurately compared with the existing regression-based models.

The remainder of this chapter is organized as follows. In Section 4.2, we provide the formulation of 2-stage SAGP model and its inference procedure. In Section 4.3, 2-stage SAGP model is evaluated experimentally using real-world data sets. Finally, this chapter is summarized and the future work is discussed in Section 4.4.

4.2 Proposed model

4.2.1 Model

In the regression approach, we distinguish between a target data set and auxiliary data sets. Let s be a target data index, and $D^t = D_s$ be the target data set; let $D^a = D \setminus D_s$ be the $S - 1$ auxiliary data sets, where $D \setminus D_s$ represents the set difference of D and D_s . In this section, we use the following notations for target and auxiliary data sets,

$$D^t = D_s = \{(R_{s,n}, y_{s,n}) \mid n = 1, \dots, |P_s|\}, \quad (4.1)$$

$$D^a = \bigcup_{s' \in S^a} D_{s'}, \text{ where } D_{s'} = \{(R_{s',n}, y_{s',n}) \mid n = 1, \dots, |P_{s'}|\}, \quad (4.2)$$

where $S^a = \{1, \dots, S\} \setminus s$ is the indices for the auxiliary data sets. Let $\mathbf{y}_s = (y_{s,1}, \dots, y_{s,|P_s|})^\top$ be a $|P_s|$ -dimensional vector consisting of the observations for the target data set. Let $\mathbf{y}_{s'} = (y_{s',1}, \dots, y_{s',|P_{s'}|})^\top$ denote a $|P_{s'}|$ -dimensional vector consisting of the observations for the s' -th auxiliary data set. In the regression approach, we aim to model a conditional probability $p(\mathbf{y}_s \mid \{\mathbf{y}_{s'}\}_{s' \in S^a})$, which allows us to adopt a two-step inference procedure described in Section 4.2.2; this is advantageous in the computational cost for learning model parameters.

We propose 2-stage SAGP model (two-stage spatially aggregated Gaussian process model) for refining coarse-grained target data $\mathbf{y}_s$ by utilizing multiple auxiliary data sets $\{\mathbf{y}_{s'}\}_{s' \in S^a}$

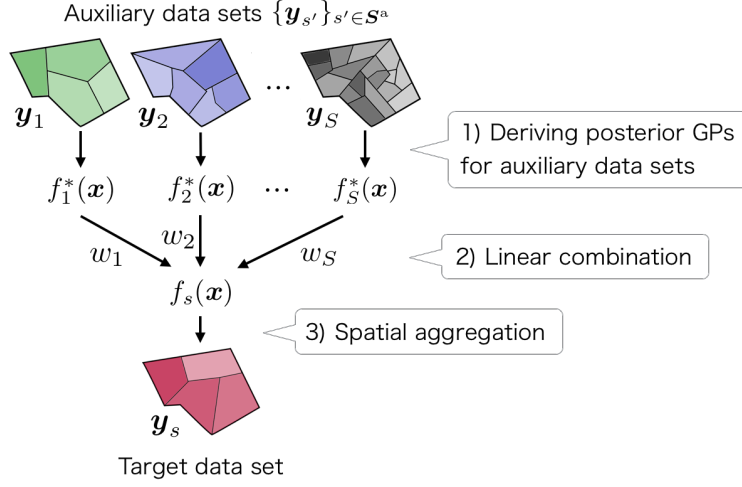


FIGURE 4.1: Generative process for 2-stage SAGP model.

with various granularities. The formulation of this model consists of three parts: 1) deriving the posterior GPs for auxiliary data sets; 2) constructing a GP for target data as a linear combination of the posterior GPs for auxiliary data sets; 3) introducing an observation model for target aggregated data. The generative process for 2-stage SAGP model is illustrated schematically in Figure 4.1. Mathematical notations for 2-stage SAGP model are summarized in Table 4.1.

Deriving the posterior GPs for auxiliary data sets

In order to handle multiple auxiliary data sets with various granularities, we first apply SAGP-S (described in Chapter 3) to each auxiliary data set, separately, to derive a posterior GP defined on the continuous space; this conceptually corresponds to spatial interpolation for aggregated data. Let $f_{s'}(x)$ be a function for the s' -th auxiliary data set. Consider $S - 1$ independent GPs,

$$f_{s'}(x) \sim \mathcal{GP}(\nu_{s'}(x), \gamma_{s'}(x, x')), \quad s' \in S^a, \quad (4.3)$$

with a mean function $\nu_{s'}(x) : \mathbf{X} \rightarrow \mathbb{R}$ and a covariance function $\gamma_{s'}(x, x') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ for the s' -th auxiliary data set, both of which are assumed integrable. We assume that the auxiliary data sets $\{y_{s'}\}_{s' \in S^a}$ are independently generated from the following observation models with spatial aggregation processes,

$$y_{s'} | f_{s'}(x) \sim \mathcal{N}\left(y_{s'} \mid \int_{\mathbf{X}} a_{s'}(x) f_{s'}(x) dx, \sigma_{s'}^2 \mathbf{I}\right), \quad s' \in S^a, \quad (4.4)$$

TABLE 4.1: Notation for 2-stage SAGP model.

Symbol	Description
s	index of target data set
$\mathbf{S}^a$	indices of auxiliary data sets, $\mathbf{S}^a = \{1, \dots, S\} \setminus s$
s'	index of auxiliary data set, $s' \in \mathbf{S}^a$
$f_{s'}(\mathbf{x})$	GP for the s' -th auxiliary data set
$\nu_{s'}(\mathbf{x})$	mean function of $f_{s'}(\mathbf{x})$
$\gamma_{s'}(\mathbf{x}, \mathbf{x}')$	covariance function of $f_{s'}(\mathbf{x})$
$\sigma_{s'}^2$	noise variance for the s' -th auxiliary data set
$w_{s'}$	regression coefficient for the s' -th auxiliary data set, $w_{s'} \in \mathbb{R}$
$f_s(\mathbf{x})$	GP for the target data set
$m_s(\mathbf{x})$	mean function of $f_s(\mathbf{x})$
$k_s(\mathbf{x}, \mathbf{x}')$	covariance function of $f_s(\mathbf{x})$
σ_s^2	noise variance for the target data set
$\eta(\mathbf{x})$	GP for the regression residuals in $f_s(\mathbf{x})$
$\nu(\mathbf{x})$	mean function of $\eta(\mathbf{x})$
$\gamma(\mathbf{x}, \mathbf{x}')$	covariance function of $\eta(\mathbf{x})$

where $\mathbf{a}_{s'}(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}^{|\mathcal{P}_{s'}|}$ is a nonnegative weight function for spatial aggregation of the s' -th auxiliary data set, and where $\sigma_{s'}^2$ is a noise variance. Defining $f_{s'}^*(\mathbf{x})$ as the posterior GP for the s' -th auxiliary data set, the posterior GPs for the auxiliary data sets are given by

$$f_{s'}^*(\mathbf{x}) \sim \mathcal{GP}(\nu_{s'}^*(\mathbf{x}), \gamma_{s'}^*(\mathbf{x}, \mathbf{x}')), \quad s' \in \mathbf{S}^a, \quad (4.5)$$

where $\nu_{s'}^*(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}$ and $\gamma_{s'}^*(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ are the mean function and the covariance function, respectively, for the s' -th posterior GP. The mean function $\nu_{s'}^*(\mathbf{x})$ and the covariance function $\gamma_{s'}^*(\mathbf{x}, \mathbf{x}')$ are given by

$$\nu_{s'}^*(\mathbf{x}) = \nu_{s'}(\mathbf{x}) + \mathbf{h}_{s'}(\mathbf{x})^\top \mathbf{C}_{s'}^{-1}(\mathbf{y}_{s'} - \boldsymbol{\mu}_{s'}), \quad (4.6)$$

$$\gamma_{s'}^*(\mathbf{x}, \mathbf{x}') = \gamma_{s'}(\mathbf{x}, \mathbf{x}') - \mathbf{h}_{s'}(\mathbf{x})^\top \mathbf{C}_{s'}^{-1} \mathbf{h}_{s'}(\mathbf{x}'), \quad (4.7)$$

respectively, where

$$\boldsymbol{\mu}_{s'} = \int_{\mathbf{X}} \mathbf{a}_{s'}(\mathbf{x}) \nu_{s'}(\mathbf{x}) d\mathbf{x}, \quad (4.8)$$

$$\mathbf{C}_{s'} = \iint_{\mathbf{X} \times \mathbf{X}} \gamma_{s'}(\mathbf{x}, \mathbf{x}') \mathbf{a}_{s'}(\mathbf{x}) \mathbf{a}_{s'}(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \sigma_{s'}^2 \mathbf{I}, \quad (4.9)$$

$$\mathbf{h}_{s'}(\mathbf{x}) = \int_{\mathbf{X}} \mathbf{a}_{s'}(\mathbf{x}') \gamma_{s'}(\mathbf{x}', \mathbf{x}) d\mathbf{x}'. \quad (4.10)$$

See Chapter 3 for more information about the derivation of the posterior GPs.

GP for target data

In 2-stage SAGP model, the function for the target data is assumed to be a GP that is modeled by a linear combination of the posterior GPs $\{f_{s'}^*(\mathbf{x})\}_{s' \in \mathcal{S}^a}$ for the auxiliary data sets. Defining $f_s(\mathbf{x})$ as the GP for target data, $f_s(\mathbf{x})$ is given by

$$f_s(\mathbf{x}) = \mathbf{w}^\top \mathbf{f}^*(\mathbf{x}) + \eta(\mathbf{x}), \quad (4.11)$$

where $\mathbf{f}^*(\mathbf{x}) = (f_1^*(\mathbf{x}), \dots, f_S^*(\mathbf{x}))^\top$, and where $\mathbf{w} = (w_1, \dots, w_S)^\top \in \mathbb{R}^{S-1}$ is a weight vector for the posterior GPs of the auxiliary data sets, and where

$$\eta(\mathbf{x}) \sim \mathcal{GP}(\nu(\mathbf{x}), \gamma(\mathbf{x}, \mathbf{x}')) \quad (4.12)$$

is a GP for the regression residuals in $f_s(\mathbf{x})$. Here, $\nu(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}$ and $\gamma(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ are the mean function and the covariance function, respectively, both of which are assumed integrable. It is known that a linear combination of GPs is again a GP; thus, $f_s(\mathbf{x})$ (4.11) is given by

$$f_s(\mathbf{x}) \sim \mathcal{GP}(m_s(\mathbf{x}), k_s(\mathbf{x}, \mathbf{x}')), \quad (4.13)$$

where the mean function $m_s(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}$ is given by

$$\begin{aligned} m_s(\mathbf{x}) &= \mathbb{E}[f_s(\mathbf{x})] \\ &= \mathbf{w}^\top \mathbb{E}[\mathbf{f}^*(\mathbf{x})] + \mathbb{E}[\eta(\mathbf{x})] \\ &= \mathbf{w}^\top \mathbf{v}^*(\mathbf{x}) + \nu(\mathbf{x}), \end{aligned} \quad (4.14)$$

and where the covariance function $k_s(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ is given by

$$\begin{aligned} k_s(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f_s(\mathbf{x}) - m_s(\mathbf{x}))(f_s(\mathbf{x}') - m_s(\mathbf{x}'))^\top] \\ &= \mathbb{E}\left[\left(\mathbf{w}^\top (\mathbf{f}^*(\mathbf{x}) - \mathbf{v}^*(\mathbf{x})) + (\eta(\mathbf{x}) - \nu(\mathbf{x}))\right) \left(\mathbf{w}^\top (\mathbf{f}^*(\mathbf{x}') - \mathbf{v}^*(\mathbf{x}')) + (\eta(\mathbf{x}') - \nu(\mathbf{x}'))\right)^\top\right] \\ &= \mathbf{w}^\top \mathbf{\Gamma}^*(\mathbf{x}, \mathbf{x}') \mathbf{w} + \gamma(\mathbf{x}, \mathbf{x}'). \end{aligned} \quad (4.15)$$

Here, $\mathbf{v}^*(\mathbf{x}) = (v_1^*(\mathbf{x}), \dots, v_{S-1}^*(\mathbf{x}))^\top$ and $\mathbf{\Gamma}^*(\mathbf{x}, \mathbf{x}') = \text{diag}(\gamma_1^*(\mathbf{x}, \mathbf{x}'), \dots, \gamma_{S-1}^*(\mathbf{x}, \mathbf{x}'))$. Equation (4.15) shows that the covariance function $k_s(\mathbf{x}, \mathbf{x}')$ of $f_s(\mathbf{x})$ is the linear combination of the covariance functions $\{\gamma_{s'}^*(\mathbf{x}, \mathbf{x}')\}$ of the posterior GPs $\{f_{s'}^*(\mathbf{x})\}$ for the auxiliary data sets, which allows for effectively learning the weights $\mathbf{w}$ while considering not only the strength of relationship with the target data but also the posterior variances of each auxiliary data set; details are given in the second inference step paragraph of Section 4.2.2.

Algorithm 1: Bayesian inference procedure of posterior GP $f^*(\mathbf{x})$ for target data.

Input : D^t, D^a **Output:** $f_s^*(\mathbf{x})$

- 1: Initialize model parameters, $\{\nu_{s'}(\mathbf{x})\}, \{\gamma_{s'}(\mathbf{x}, \mathbf{x})\}, \{\sigma_{s'}^2\}, \nu(\mathbf{x}), \gamma(\mathbf{x}, \mathbf{x}), \sigma_s^2, \mathbf{w}$
 - 2: /* first inference step */
 - 3: **for** $s' \in S^a$ **do**
 - 4: Estimate $\nu_{s'}(\mathbf{x}), \gamma_{s'}(\mathbf{x}, \mathbf{x}), \sigma_{s'}^2$ by maximizing the logarithm of (4.17)
 - 5: **end for**
 - 6: /* second inference step */
 - 7: Estimate $\mathbf{w}, \nu(\mathbf{x}), \gamma(\mathbf{x}, \mathbf{x}), \sigma_s^2$ by maximizing the logarithm of (4.18)
 - 8: Construct posterior GP by (4.25) using the estimated model parameters
-

Observation model for target data

In a way similar to SAGP-S, the target observations $\mathbf{y}_s$ are assumed to be obtained by integrating the GP $f_s(\mathbf{x})$ over the corresponding regions. The observation model of $\mathbf{y}_s$ is given by ¹

$$\mathbf{y}_s \mid f_s(\mathbf{x}) \sim \mathcal{N} \left(\mathbf{y}_s \mid \int_X \mathbf{a}_s(\mathbf{x}) f_s(\mathbf{x}) d\mathbf{x}, \sigma_s^2 \mathbf{I} \right), \quad (4.16)$$

where $\mathbf{a}_s(\mathbf{x}) : X \rightarrow \mathbb{R}^{|P_s|}$ is a nonnegative weight function for target data aggregation, and where σ_s^2 is a noise variance for the target data.

4.2.2 Inference

Given the target data set D^t and the auxiliary data sets D^a , we aim to obtain the posterior GP of $f_s(\mathbf{x})$ via a Bayesian inference procedure, where the model parameters are estimated by maximizing the marginal likelihood. The estimation procedure of 2-stage SAGP model can be divided into two steps: 1) estimate parameters $\nu_{s'}(\mathbf{x}), \gamma_{s'}(\mathbf{x}, \mathbf{x}), \sigma_{s'}^2$ for each auxiliary data set, and 2) estimate regression coefficients $\mathbf{w}$ and the parameters $\nu(\mathbf{x}), \gamma(\mathbf{x}, \mathbf{x}), \sigma_s^2$ for the target data. Details of the inference procedure are shown in Algorithm 1. The parameter inference of 2-stage SAGP model can be conducted for each data set independently; thus it is advantageous in the computational cost for learning model parameters.

¹We assume that the integral appearing in (4.16) is well-defined. The assumption of the integrability of $\nu(\mathbf{x})$ and $\{\nu_{s'}(\mathbf{x})\}$ assures integrability of the mean function $m_s(\mathbf{x})$ (4.14), which allows us to discuss the conditions for the observation model (4.16) to be well-defined in a way similar to that in Appendix A.

The first inference step

Given the s' -th auxiliary data set $D_{s'}$, the marginal likelihood of $y_{s'}$ is given by

$$p(y_{s'} | v_{s'}(x), \gamma_{s'}(x, x), \sigma_{s'}^2) = \mathcal{N}(y_{s'} | \mu_{s'}, C_{s'}), \quad (4.17)$$

where $\mu_{s'}$ and $C_{s'}$ are (4.6) and (4.7), respectively. The parameters $v_{s'}(x), \gamma_{s'}(x, x), \sigma_{s'}^2$ are estimated by maximizing the logarithm of (4.17). By solving the optimization problem for each auxiliary data set independently, we obtain the set of the estimated parameters for all auxiliary data sets. The posterior GPs $\{f_{s'}^*(x)\}$ corresponding to (4.5) are obtained using the estimated parameters.

The second inference step

Given the target data set D^t , the marginal likelihood of y_s is given by

$$p(y_s | w, v(x), \gamma(x, x), \sigma_s^2) = \mathcal{N}(y_s | \mu_s, C_s). \quad (4.18)$$

Here, μ_s is a $|P_s|$ -dimensional mean vector represented by

$$\begin{aligned} \mu_s &= \int_X a_s(x) m_s(x) dx \\ &= \int_X a_s(x) (w^\top v^*(x) + v(x)) dx \\ &= \sum_{s' \in S^a} w_{s'} \int_X a_s(x) v_{s'}^*(x) dx + \int_X a_s(x) v(x) dx \\ &=: \sum_{s' \in S^a} w_{s'} \bar{v}_{s'}^* + \bar{v} \end{aligned} \quad (4.19)$$

where we use the relation (4.14), and where we define $\bar{v}_{s'}^* \in \mathbb{R}^{|P_s|}$ and $\bar{v} \in \mathbb{R}^{|P_s|}$ as

$$\bar{v}_{s'}^* = \int_X a_s(x) v_{s'}^*(x) dx, \quad (4.20)$$

$$\bar{v} = \int_X a_s(x) v(x) dx, \quad (4.21)$$

respectively. Equation (4.19) shows that the means μ_s for the target data are represented by the linear combination of values that are obtained by aggregating the posterior mean functions $\{v_{s'}^*(x)\}$ over the coarse-grained partition P_s of the target data. C_s in (4.18) is a

$|\mathbf{P}_s| \times |\mathbf{P}_s|$ covariance matrix represented by

$$\begin{aligned}
\mathbf{C}_s &= \iint_{\mathbf{X} \times \mathbf{X}} k_s(\mathbf{x}, \mathbf{x}') \mathbf{a}_s(\mathbf{x}) \mathbf{a}_s(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \sigma_s^2 \mathbf{I} \\
&= \iint_{\mathbf{X} \times \mathbf{X}} \left(\mathbf{w}^\top \mathbf{\Gamma}^*(\mathbf{x}, \mathbf{x}') \mathbf{w} + \gamma(\mathbf{x}, \mathbf{x}') \right) \mathbf{a}_s(\mathbf{x}) \mathbf{a}_s(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \sigma_s^2 \mathbf{I} \\
&= \sum_{s' \in \mathbf{S}^a} w_{s'}^2 \iint_{\mathbf{X} \times \mathbf{X}} \gamma_{s'}^*(\mathbf{x}, \mathbf{x}') \mathbf{a}_s(\mathbf{x}) \mathbf{a}_s(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \iint_{\mathbf{X} \times \mathbf{X}} \gamma(\mathbf{x}, \mathbf{x}') \mathbf{a}_s(\mathbf{x}) \mathbf{a}_s(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \sigma_s^2 \mathbf{I} \\
&=: \sum_{s' \in \mathbf{S}^a} w_{s'}^2 \bar{\mathbf{\Gamma}}_{s'}^* + \bar{\mathbf{\Gamma}} + \sigma_s^2 \mathbf{I},
\end{aligned} \tag{4.22}$$

where we use the relation (4.15), and where we define $\bar{\mathbf{\Gamma}}_{s'}^* \in \mathbb{R}^{|\mathbf{P}_s| \times |\mathbf{P}_s|}$ and $\bar{\mathbf{\Gamma}} \in \mathbb{R}^{|\mathbf{P}_s| \times |\mathbf{P}_s|}$ as

$$\bar{\mathbf{\Gamma}}_{s'}^* = \iint_{\mathbf{X} \times \mathbf{X}} \gamma_{s'}^*(\mathbf{x}, \mathbf{x}') \mathbf{a}_s(\mathbf{x}) \mathbf{a}_s(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}', \tag{4.23}$$

$$\bar{\mathbf{\Gamma}} = \iint_{\mathbf{X} \times \mathbf{X}} \gamma(\mathbf{x}, \mathbf{x}') \mathbf{a}_s(\mathbf{x}) \mathbf{a}_s(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}', \tag{4.24}$$

respectively. Equation (4.22) shows that the covariance matrix $\mathbf{C}_s$ for the target data is determined by the linear combination of the covariance matrices that are calculated via aggregating the posterior covariance functions $\{\gamma_{s'}(\mathbf{x}, \mathbf{x}')\}$ over the coarse-grained partition $\mathbf{P}_s$ of the target data.

By maximizing the logarithm of (4.18), we can obtain the model parameters. As shown in equations (4.19) and (4.22), our model can learn the regression coefficients $\mathbf{w}$ considering the strength of relationship with the target data and the posterior covariance matrix for each auxiliary data set, simultaneously. This is beneficial for determining the usefulness of each auxiliary data set in such a case that spatial interpolation (i.e., the first inference step) might fail for an auxiliary data set due to some reasons including data sparsity and non-stationarity; the weight (i.e., usefulness) for that auxiliary data set can be under-estimated by considering the posterior variances for the auxiliary data set.

Posterior GP for target data

Given the target data set $\mathbf{D}^t$, the auxiliary data sets $\mathbf{D}^a$, and the estimated parameters, the posterior GP $f_s^*(\mathbf{x})$ is given by

$$f_s^*(\mathbf{x}) \sim \mathcal{GP}(m_s^*(\mathbf{x}), k_s^*(\mathbf{x}, \mathbf{x}')), \tag{4.25}$$

where $m_s^*(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}$ and $k_s^*(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ are the mean function and the covariance function, respectively. The mean function $m_s^*(\mathbf{x})$ and the covariance function $k_s^*(\mathbf{x}, \mathbf{x}')$ are

given by

$$m_s^*(\mathbf{x}) = m_s(\mathbf{x}) + \mathbf{h}_s(\mathbf{x})^\top \mathbf{C}_s^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_s), \quad (4.26)$$

$$k_s^*(\mathbf{x}, \mathbf{x}') = k_s(\mathbf{x}, \mathbf{x}') - \mathbf{h}_s(\mathbf{x})^\top \mathbf{C}_s^{-1} \mathbf{h}_s(\mathbf{x}'), \quad (4.27)$$

respectively, where

$$\mathbf{h}_s(\mathbf{x}) = \int_{\mathbf{X}} \mathbf{a}_s(\mathbf{x}') k_s(\mathbf{x}', \mathbf{x}) d\mathbf{x}'. \quad (4.28)$$

One can obtain the posterior GP (4.25) in a fashion similar to the derivation procedure in SAGP-S. The posterior mean (4.26) can be used for predicting the data value at any location point $\mathbf{x}$; and its prediction uncertainty can be evaluated by (4.27). The prediction of $y_{s,n}^{\text{fine}}$ associated with region $\mathbf{R}_n^{\text{fine}}$ is calculated via integrating the posterior mean (4.26) over the corresponding region; one can also obtain the prediction variance at the region.

4.3 Experiments

4.3.1 Data

We conducted experiments on real-world data sets from urban cities to demonstrate the effectiveness of the proposed model. We used the data sets from two cities: New York City and Chicago, which are available on NYC Open Data (*NYC Open Data*) and Chicago Data Portal (*Chicago Data Portal*), respectively. These data sets are collected and released for improving city environments, and consist of a variety of categories including social indicators, land use, and air quality. Details of the data sets we used in the experiments are listed in Table 4.2. The number of data sets in New York City and Chicago are 10 and 3, respectively. Each data set is associated with one of the predefined geographical partitions including zip code and police precinct. The number of partition types in New York City and Chicago are 4 and 2, respectively. Table 4.2 shows the respective partition names and the number of regions in the corresponding partition. These data sets were gathered once a year at the time ranges shown in Table 4.2; the values of data were divided by the number of observation times. Data normalization was performed so that each variable in each city has zero mean and unit variance.

4.3.2 Refinement task

To evaluate the performance in refining coarse-grained aggregated data, we consider the task of predicting the fine-grained data from its coarser version: We first picked up one

TABLE 4.2: Real-world aggregated data sets.

(a) New York City			
Data	Partition	#regions	Time range
PM2.5	UHF42	42	2009 – 2010
Poverty rate	Community district	59	2009 – 2013
Unemployment rate	Community district	59	2009 – 2013
Mean commute	Community district	59	2009 – 2013
Population	Community district	59	2009 – 2013
Recycle diversion rate	Community district	59	2009 – 2013
Crime	Police precinct	77	2010 – 2016
Fire incident	Zip code	186	2010 – 2016
311 call	Zip code	186	2010 – 2016
Public telephone	Zip code	186	2016

(b) Chicago			
Data	Partition	#regions	Time range
Crime	Police precinct	25	2012
Poverty rate	Community district	77	2008 – 2012
Unemployment rate	Community district	77	2008 – 2012

target data set and used its coarser version to learn model parameters; then we predicted the original fine-grained target data by using the learned model. It should be noted that we did not use the fine-grained target data in the learning phase. In the experiments, every data set, except for Crime data in Chicago, was used as a target data set; the coarser version of Crime data in Chicago for training is not available online. The coarse-grained partitions for New York City and Chicago are Borough (5) and Side (9), respectively, where each number in parentheses denotes the number of regions contained in the corresponding coarse-grained partition.

The evaluation metric is the mean absolute percentage error (MAPE) of the fine-grained target values,

$$\frac{1}{|\mathcal{P}^{\text{fine}}|} \sum_{n=1}^{|\mathcal{P}^{\text{fine}}|} \left| \frac{(y_{s,n}^* - y_{s,n}^{\text{fine}})}{y_{s,n}^{\text{fine}}} \right|, \quad (4.29)$$

where $y_{s,n}^{\text{fine}}$ is the true value and $y_{s,n}^*$ is its predicted value.

4.3.3 Setup of 2-stage SAGP model

In our experiments, all GPs in 2-stage SAGP model are assumed to be zero-mean GPs. We used the following squared-exponential kernel as the covariance function for all GPs,

$$\gamma_{s'}(\mathbf{x}, \mathbf{x}') = \alpha_{s'}^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\beta_{s'}^2}\right), \quad s' \in \mathbf{S}^a, \quad (4.30)$$

$$\gamma_s(\mathbf{x}, \mathbf{x}') = \alpha_s^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\beta_s^2}\right), \quad (4.31)$$

$$\gamma(\mathbf{x}, \mathbf{x}') = \alpha^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\beta^2}\right), \quad (4.32)$$

where $\{\alpha_{s'}^2\}, \alpha_s^2, \alpha^2$ are signal variances that control the magnitude of the covariance, and where $\{\beta_{s'}\}, \beta_s, \beta$ are scale parameters that determine the degrees of spatial correlation. Here, $\|\cdot\|$ is the Euclidean norm. The model parameters are estimated by minimizing the negative log marginal likelihood (4.17) and (4.18); we solve the optimization problems using the L-BFGS method (Liu and Nocedal, 1989). We used the $300 \text{ m} \times 300 \text{ m}$ grid cells to approximate the integral over regions (see Section 3.3.4); the resulting sets $\mathbf{G}$ of grid points for New York City and Chicago consisted of 9,352 and 7,400 grid points, respectively.

4.3.4 Baselines

We compared 2-stage SAGP model with the naive Gaussian process regression (GPR) (Rasmussen and Williams, 2006), SAGP-S, linear regression model (LR) (Smith, Mashhadi, and Capra, 2014), and two-stage statistical downscaling method (2-stage SD) (Park, 2013). In GPR and SAGP-S, the fine-grained target data were predicted using only the coarse-grained target data: They did not use auxiliary data sets. LR and 2-stage SD require that the spatial granularities of all auxiliary data sets are the same as the desired granularity of target data. The auxiliary data sets, however, have various granularities; thus SAGP-S was first applied so that the granularities of the auxiliary data sets match with that of the fine-grained partition, and used the predictions of auxiliary data at the desired fine-grained regions. Let $f_{s',n}^*$ denote the prediction at the n -th region of the s' -th auxiliary data set, which is obtained by integrating the posterior mean $\nu_{s'}^*(\mathbf{x})$ (4.6) of $f_{s'}^*(\mathbf{x})$ over the region $\mathbf{R}_n^{\text{fine}}$, represented by

$$f_{s',n}^* = \int_{\mathbf{X}} a_n^{\text{fine}}(\mathbf{x}) \nu_{s'}^*(\mathbf{x}) d\mathbf{x}, \quad (4.33)$$

where

$$a_n^{\text{fine}}(\mathbf{x}) = \frac{\mathbb{1}(\mathbf{x} \in \mathbf{R}_n^{\text{fine}})}{\int_{\mathbf{X}} \mathbb{1}(\mathbf{x}' \in \mathbf{R}_n^{\text{fine}}) d\mathbf{x}'} \cdot \quad (4.34)$$

Descriptions of baselines are as follows.

- **GPR.** A simple spatial interpolation was applied to the coarse-grained target data. The standard GPR assumes that each sample is observed at a location point. We thus regarded that each observation is associated with the centroid of each region.
- **SAGP-S.** We applied SAGP-S (described in Chapter 3) to the coarse-grained target data. Different from the standard GPR, it can handle the observations associated with regions via the observation model with spatial aggregation processes.
- **LR.** A simple linear regression has been used for analyzing the relationships between target and auxiliary data sets (Bogomolov et al., 2014; Smith, Mashhadi, and Capra, 2014). In the training phase, all auxiliary data sets are first aggregated into the coarse-grained partition of target data via spatial averaging; the regression coefficients for each auxiliary data set are then estimated. The predictions of fine-grained target data are obtained by applying the learned regression model to the refined auxiliary data $\{f_{s',n}^*\}$ (4.33). The predicted value $y_{s,n}^*$ is then given by

$$y_{s,n}^* = \hat{w}_0 + \sum_{s' \in \mathcal{S}^a} \hat{w}_{s'} f_{s',n}^*. \quad (4.35)$$

Here $\hat{w}_0$ is the estimated bias parameter, and $\hat{w}_{s'}$ is the estimated regression coefficient for the s' -th auxiliary data set.

- **2-stage SD.** In 2-stage SD, the target data are assumed to be expressed by a sum of linear regression terms and residual terms. Unlike LR, 2-stage SD can predict the fine-grained target data while encouraging the consistency between coarse- and fine-grained target data. The inference procedure of 2-stage SD can be divided into two stages. In the first stage, the regression coefficients are estimated in a way similar to the training process of LR. In the second stage, given the estimated coefficients $\hat{w}_0, \{\hat{w}_{s'}\}$ and the refined auxiliary data sets $\{f_{s',n}^*\}$, the fine-grained target data $\{y_{s,n}^*\}$ are estimated to be those that satisfy the following relation:

$$\begin{aligned} y_{s,n'} &= \underbrace{\hat{w}_0 + \sum_{s' \in \mathcal{S}^a} \hat{w}_{s'} \left[\frac{1}{|\mathcal{P}_{n'}^{\text{fine}}|} \sum_{n=1}^{|\mathcal{P}_{n'}^{\text{fine}}|} f_{s',n}^* \right]}_{\text{linear regression term}} + \underbrace{r_{n'}^{\text{coar}}}_{\text{residual term}} \\ &= \frac{1}{|\mathcal{P}_{n'}^{\text{fine}}|} \sum_{n=1}^{|\mathcal{P}_{n'}^{\text{fine}}|} \left[\hat{w}_0 + \sum_{s' \in \mathcal{S}^a} \hat{w}_{s'} f_{s',n}^* + r_n^{\text{fine}} \right] \\ &= \frac{1}{|\mathcal{P}_{n'}^{\text{fine}}|} \sum_{n=1}^{|\mathcal{P}_{n'}^{\text{fine}}|} y_{s,n'}^* \end{aligned} \quad (4.36)$$

where $\mathcal{P}_{n'}^{\text{fine}}$ is a set of fine-grained regions contained in the n' -th coarse-grained

region $R_{s,n'}$, and where $r_{n'}^{\text{coar}}$ and r_n^{fine} are regression residuals for coarse- and fine-grained target data, respectively. The relation (4.36) expresses the spatial aggregation constraints, which states that the value $y_{s,n'}$ at the coarse-grained region $R_{s,n'}$ is the average of the values $\{y_{s,n}^*\}$ associated with the fine-grained regions in $P_{n'}^{\text{fine}}$. The linear regression terms in (4.36) have already been fixed in the first stage; thus the residual term $r_{n'}^{\text{coar}}$ can be determined. The residual r_n^{fine} is obtained by spatial interpolation method, e.g., block kriging (Burgess and Webster, 1980), given the set of coarse-grained residuals $\{r_{n'}^{\text{coar}}\}$.

4.3.5 Results

Refinement performance

Table 4.3 shows MAPE and standard errors for GPR, SAGP-S, LR, 2-stage SD, and 2-stage SAGP model. Compared SAGP-S with GPR, the errors of SAGP-S were lower than those of GPR for all data sets. These results show that the incorporation of aggregation constraints is essential for modeling data aggregated over regions. For all data sets, 2-stage SAGP model achieved better performance than the other methods; we evaluated the differences between 2-stage SAGP model and the baselines using Student's t-test (See Table 4.3). These results indicate that 2-stage SAGP model can utilize multiple aggregated data sets with various granularities to accurately refine coarse-grained target data compared with the existing regression-based models.

Figures 4.2 and 4.3 visualize the refined target data for poverty rate data and recycle diversion rate data, respectively, in New York City. Figure 4.4 shows the refined target data for poverty rate data in Chicago. For each data set, we illustrate the true fine-grained target data on the left, and the refined target data for 2-stage SAGP model, 2-stage SD, LR, SAGP-S, and GPR on the right, where darker hues represent regions with higher values; the values of each method were normalized to the range $[0, 1]$. As shown, one observes that 2-stage SAGP model can better discern the fine-grained distribution of each data set.

TABLE 4.3: Performance comparisons for 2-stage SAGP model: MAPE and standard errors for the prediction of fine-grained target data in New York City and Chicago. The single star (*) and the double star (**) indicate significant difference between 2-stage SAGP model and other models at the levels of P values of < 0.05 and < 0.01 , respectively.

(a) New York City					
	GPR	SAGP-S	LR	2-stage SD	2-stage SAGP model
PM2.5	0.072 ± 0.010	0.053 ± 0.007	0.046 ± 0.006	0.045 ± 0.005	$0.042 \pm 0.005^{**}$
Poverty rate	0.344 ± 0.046	0.281 ± 0.034	0.266 ± 0.029	0.247 ± 0.022	$0.218 \pm 0.021^{**}$
Unemployment rate	0.319 ± 0.036	0.232 ± 0.025	0.200 ± 0.021	0.187 ± 0.018	$0.154 \pm 0.021^*$
Mean commute	0.131 ± 0.020	0.076 ± 0.010	0.072 ± 0.010	0.067 ± 0.008	$0.052 \pm 0.005^*$
Population	0.577 ± 0.104	0.371 ± 0.040	0.359 ± 0.030	0.353 ± 0.034	0.321 ± 0.039
Recycle diversion rate	0.353 ± 0.049	0.282 ± 0.034	0.263 ± 0.035	0.260 ± 0.031	$0.191 \pm 0.023^{**}$
Crime	0.860 ± 0.102	0.525 ± 0.112	0.491 ± 0.111	0.479 ± 0.107	$0.413 \pm 0.086^*$
Fire incident	1.097 ± 0.097	0.793 ± 0.148	0.627 ± 0.114	0.631 ± 0.134	$0.560 \pm 0.106^{**}$
311 call	0.083 ± 0.004	0.069 ± 0.005	0.066 ± 0.004	0.066 ± 0.004	$0.061 \pm 0.004^*$
Public telephone	0.131 ± 0.008	0.099 ± 0.008	0.095 ± 0.008	0.099 ± 0.008	$0.089 \pm 0.008^{**}$
(b) Chicago					
	GPR	SAGP-S	LR	2-stage SD	2-stage SAGP model
Poverty rate	0.599 ± 0.099	0.423 ± 0.063	0.387 ± 0.065	0.363 ± 0.050	$0.274 \pm 0.035^{**}$
Unemployment rate	0.478 ± 0.047	0.405 ± 0.041	0.323 ± 0.037	0.297 ± 0.031	$0.257 \pm 0.030^*$

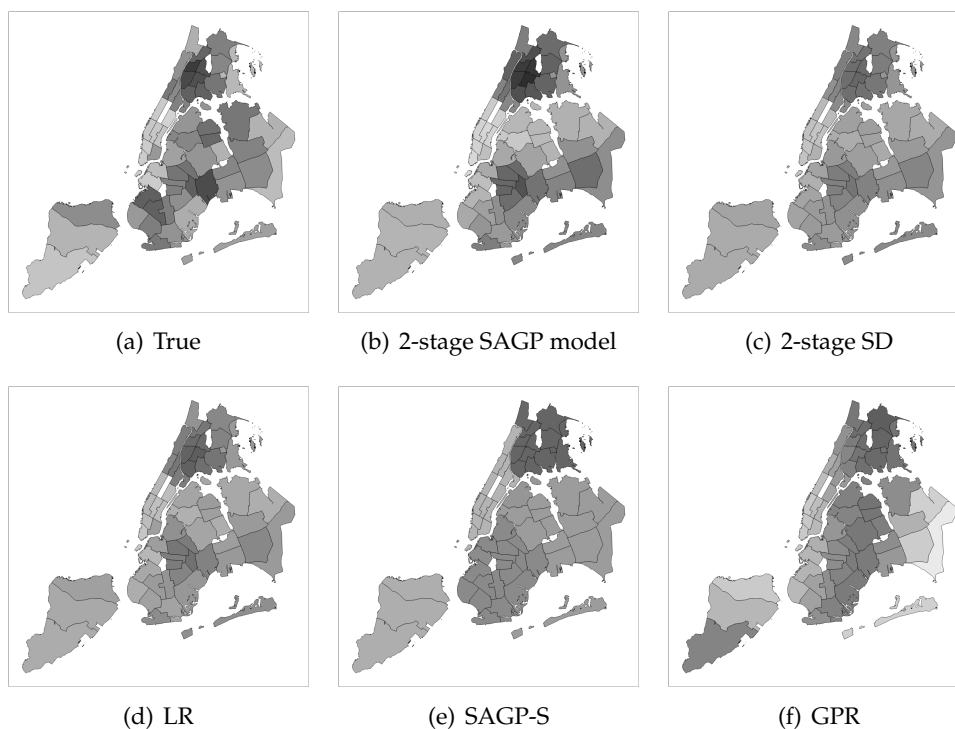


FIGURE 4.2: Comparison of the refined poverty rate data in New York City (regression approaches).

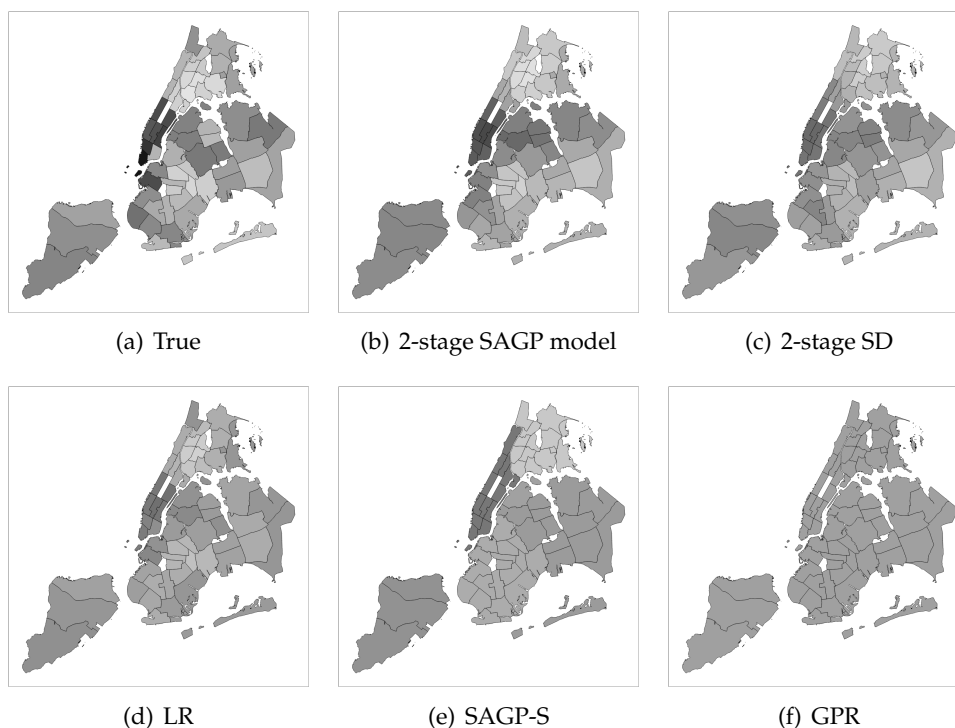


FIGURE 4.3: Comparison of the refined recycle diversion rate data in New York City (regression approaches).

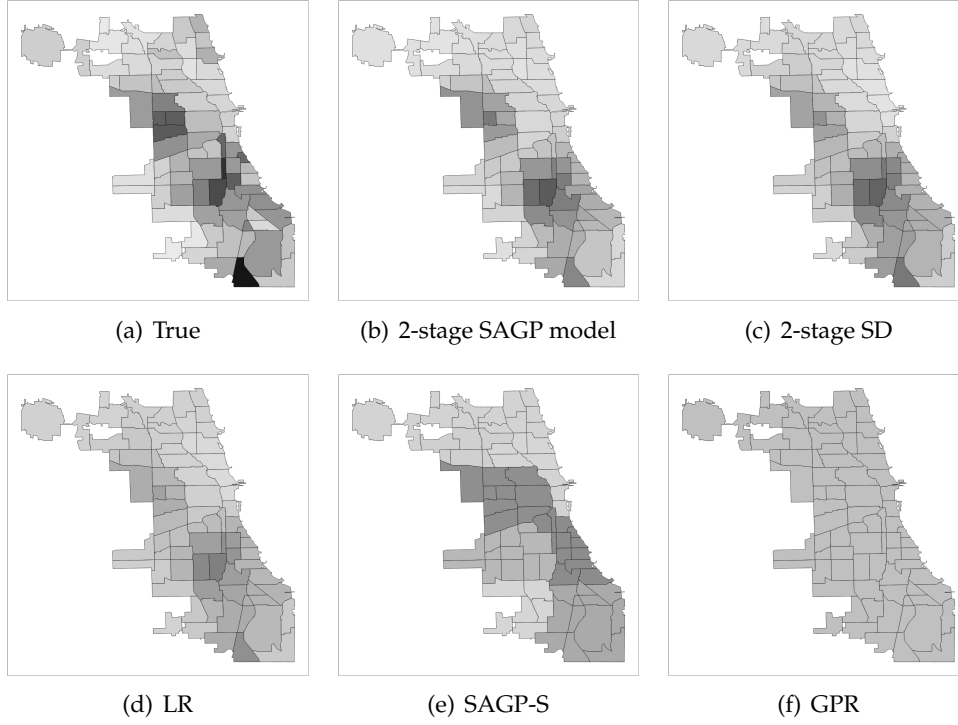


FIGURE 4.4: Comparison of the refined poverty rate data in Chicago (regression approaches).

Evaluation of auxiliary aggregated data sets

We conducted exploratory analysis to demonstrate the effectiveness of 2-stage SAGP model. The key characteristic of our model is to learn the regression coefficients while considering the prediction variances for each auxiliary data set (described in (4.18)); this is the major difference from the existing methods (i.e., 2-stage SD). In the following, we demonstrate that it is effective for evaluating the usefulness of each auxiliary data set.

Tables 4.4(a) and 4.4(b) show the rankings of auxiliary data sets when we used the poverty rate data and the recycle diversion rate data in New York City, respectively, as the target data set. The auxiliary data sets in these tables were arranged in descending order of the absolute values $|w_{s'}|$ of the estimated regression coefficients. The average of prediction variances for each auxiliary data set was listed in the *Variance* columns in each table; for the s' -th auxiliary data set, the average of variances at regions is given by

$$\frac{1}{|P_s|} \sum_{n=1}^{|P_s|} \bar{\Gamma}_{s'}^*(n, n), \quad (4.37)$$

where $\bar{\Gamma}_{s'}^*(n, n)$ is the (n, n) -entry of $\bar{\Gamma}_{s'}^*$ (4.23). Let us focus on Crime and Population data sets (illustrated in bold letters in Tables 4.4(a) and 4.4(b)). One can confirm that 2-stage

TABLE 4.4: Ranking of auxiliary data sets based on the absolute values $|w_{s'}|$ of the regression coefficients estimated by 2-stage SAGP model and 2-stage SD.

(a) Ranking when using poverty rate data in New York City as target data set.					
2-stage SAGP model			2-stage SD		
Auxiliary data set	$w_{s'}$	Variance	Auxiliary data set	$w_{s'}$	
1. Unemployment rate	1.237	0.016	Unemployment rate	1.142	
2. 311 call	0.556	0.015	Crime	-1.129	
3. Crime	-0.524	0.055	Population	-0.694	
4. Fire incident	0.463	0.010	PM2.5	0.543	
5. Mean commute	0.406	0.012	Mean commute	0.479	
6. Population	-0.266	0.045	311 call	0.436	
7. Recycle diversion rate	-0.259	0.009	Fire incident	0.344	
8. Public telephone	-0.147	0.012	Public telephone	0.265	
9. PM2.5	0.044	0.031	Recycle diversion rate	-0.038	

(b) Ranking when using recycle diversion rate data in New York City as target data set.					
2-stage SAGP model			2-stage SD		
Auxiliary data set	$w_{s'}$	Variance	Auxiliary data set	$w_{s'}$	
1. Unemployment rate	-1.359	0.016	Unemployment rate	-1.204	
2. 311 call	-0.564	0.015	Crime	0.717	
3. Public telephone	0.492	0.012	Population	0.539	
4. Fire incident	-0.391	0.010	Mean commute	-0.417	
5. Crime	0.376	0.055	PM2.5	-0.401	
6. Mean commute	-0.321	0.012	Poverty rate	0.367	
7. Population	0.224	0.045	311 call	-0.276	
8. PM2.5	0.162	0.031	Fire incident	-0.168	
9. Poverty rate	0.133	0.019	Public telephone	-0.054	

SAGP model assigned relatively small coefficients to these auxiliary data sets, compared with 2-stage SD. As shown in the *Variance* columns, one observes that the variances for these data sets were larger than those for the other data sets. These results indicate that 2-stage SAGP model can discern the usefulness of each auxiliary data set considering not only the strength of relationships with the target data but also the prediction uncertainty in spatial interpolation of auxiliary data sets.

Figure 4.5 shows the visualization results of the prediction variances for the upper-ranked data sets in Tables 4.4(a) and 4.4(b). The left figures in Figure 4.5 represent the input data $y_{s'}$; the middle figures are the means $\bar{v}_{s'}^*$ (4.19) for each data set; the right figures represent

the (n, n) -entry $\bar{\Gamma}_{s'}^*(n, n)$ of the covariance matrix $\bar{\Gamma}_{s'}^*$ (4.23) for each data set. One observes that the variances at the regions located at the edge of the city tend to have larger values compared with those inside the city. This is reasonable because extrapolation is generally more difficult than interpolation. As shown in Figure 4.5, the prediction variances for Crime and Population data sets were actually larger than those of the other data sets. In Crime data set, the input data has highly non-stationary property (illustrated in the left figure in Figure 4.5); thus, its variance might be larger. These results show that 2-stage SAGP model is robust even if the spatial interpolation for auxiliary data sets is difficult due to some reasons involved in data properties (e.g., non-stationarity and data sparsity).

4.4 Discussion

In this chapter, we proposed a new regression-based model, two-stage spatially aggregated Gaussian process model (2-stage SAGP model), that can refine coarse-grained aggregated data by utilizing auxiliary data sets with various granularities. We also developed the inference procedure based on Bayesian approach. With the proposed model, the Gaussian process (GP) for target data is assumed to be a linear combination of the posterior GPs for auxiliary data sets. We adopt an observation model with aggregation processes to handle data that have been aggregated over regions. The parameter estimation procedure can be divided into two steps, thus it has the advantage in the computational efficiency. We will discuss the computation time of inference processes in Section 5.3.

In the experiments, we confirmed that 2-stage SAGP model achieves better performance for coarse-grained data refinement compared with existing regression-based models. We experimentally analyzed the relation between regression coefficients and posterior variances for auxiliary data sets; these results show that our model can evaluate the usefulness of each auxiliary data set by considering not only the strength of relationships with target data but also the uncertainty in spatial interpolation of auxiliary data.

One drawback of 2-stage SAGP model is that it could not make full use of the auxiliary data set that have coarse granularity. This is because the posterior GP derivation is separately conducted for each auxiliary data set; this model might fail to estimate the spatial correlation precisely due to data sparsity issues. A promising approach to address these issues is the use of multivariate modeling, which will be described in Chapter 5.

One of the future works is to incorporate rich information on the map, e.g., satellite images. Some works have experimentally shown that the satellite images are useful for performance improvements; the method developed in (Law et al., 2018) adopted deep neural networks for handling unstructured high-dimensional data (i.e., images).

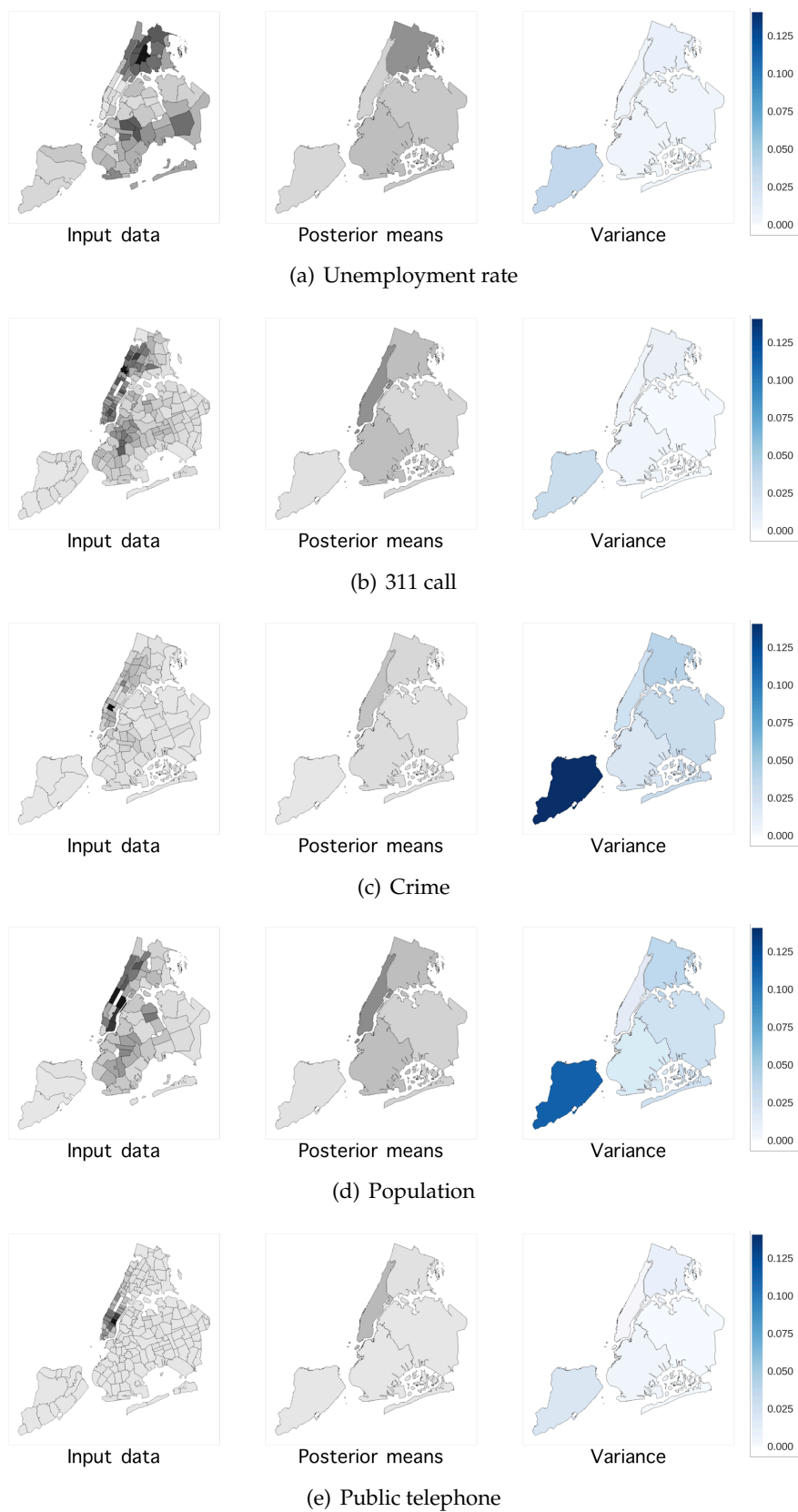


FIGURE 4.5: Visualization of the prediction variances for auxiliary data sets in New York City.

Chapter 5

Multivariate approach

5.1 Overview

In the previous chapter, we describe the regression-based method, i.e., 2-stage SAGP model, to use multiple aggregated data sets with various granularities for the data refinement. In 2-stage SAGP model, the spatial interpolation is separately conducted for each auxiliary data set, which provides the efficient parameter estimation. However, this makes it difficult to accurately interpolate the coarse-grained auxiliary data due to the data sparsity issues. In other words, 2-stage SAGP model implicitly requires that the spatial granularities of all data sets, except for target data, are fine enough to appropriately interpolate the data. That is why 2-stage SAGP model may not make full use of multiple aggregated data sets, some of which have coarse granularity.

In this chapter, we discuss the methods based on the multivariate approach, in which variables for all data sets are assumed to follow a joint distribution of them. Multivariate modeling is generally advantageous in that it can avoid the overfitting to sparse data by taking the help of dependences between data sets. As described in Section 2.2.3, the major limitation of existing multivariate models is that data samples are needed to be observed at location points, namely these models are applicable only for individual-level data.

We propose a probabilistic model, *Spatially Aggregated Gaussian Processes with Multiple Outputs (SAGP-M)*, that can infer multivariate function from multiple aggregated data sets with various granularities. SAGP-M is a natural extension of SAGP-S for handling a multivariate random variable. In SAGP-M, the functions for the data sets are assumed to be a multivariate dependent GP that is modeled as a linear mixing of independent latent GPs. Sharing the latent GPs among all data sets enables us to effectively learn the spatial correlation for each data set even if the numbers of observations in some data sets are small; that is, some data sets are associated with coarse-grained partitions. Since the aggregated data are identified by regions, not by location points, we design an observation model

with the spatial aggregation process, as in SAGP-S, in which multivariate observations are assumed to be calculated by integrating the mixed GP over the corresponding region.

The mechanism adopted in SAGP-M for sharing latent processes is also advantageous in that it makes it straightforward to utilize data sets from multiple domains (e.g., cities). This allows our model to learn the spatial correlation for each data set by sharing the latent GPs among all data sets from multiple domains; SAGP-M remains applicable even if we have only a few data sets available for a single domain.

The inference of SAGP-M is based on a Bayesian inference procedure. The model parameters can be estimated by maximizing the marginal likelihood, in which all the GPs are still analytically integrated out. Different from 2-stage SAGP model, SAGP-M learns all model parameters simultaneously. By deriving the posterior GP, we can predict the data value at any location point considering the spatial correlations and the dependences between aggregated data sets, simultaneously.

We conduct experiments on multiple real-world data sets from New York City and Chicago. The results show that our model can accurately refine coarse-grained aggregated data, and improve refinement performances by utilizing data sets from multiple cities.

The remainder of this chapter is organized as follows. Section 5.2 presents the formulation of SAGP-M and its inference procedure. In Section 5.3, the effectiveness of SAGP-M is demonstrated using real-world data sets. Finally, we summarize this chapter and discuss the future work in Section 5.4.

5.2 Proposed model

5.2.1 Model

In the multivariate approach, we do not distinguish between target and auxiliary data sets; the notations for aggregated data are shown in Table 2.1. Let $\mathbf{y}_s = (y_{s,1}, \dots, y_{s,|P_s|})$ be a $|P_s|$ -dimensional vector consisting of the observations for the s -th aggregated data set. Let $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_S)^\top$ denote an N -dimensional vector concatenating the observations for all aggregated data sets, where $N = \sum_{s=1}^S |P_s|$ is the total number of observations. In the multivariate approach, we attempt to model a joint probability $p(\mathbf{y})$; thus the model parameters for each data set can be learned with consideration of dependences between data sets.

We propose SAGP-M (spatially aggregated Gaussian processes with multiple outputs) for inferring the multivariate function from multiple aggregated data sets with various granularities. In the following, let us first consider a formulation in the case of a single domain

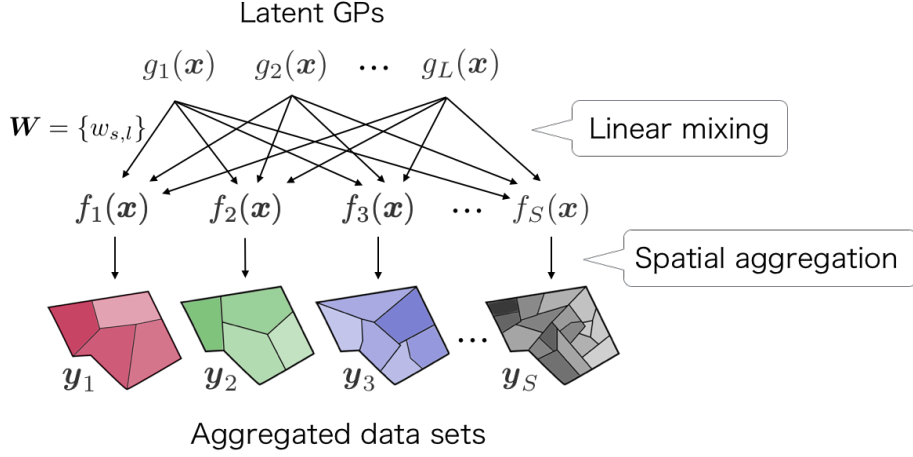


FIGURE 5.1: Generative process for SAGP-M.

(e.g., a city), then we mention an extension to the case of multiple domains. Figure 5.1 schematically illustrates the generative process for SAGP-M in the case of a single domain. Mathematical notations for SAGP-M in the case of a single domain are summarized in Table 5.1

Formulation for the case of a single domain

In SAGP-M, the functions for the respective data sets on the continuous space are assumed to be the dependent Gaussian process (GP) with multiple outputs. The multivariate dependent GP is first constructed by linearly mixing some independent latent GPs. Consider L independent GPs,

$$g_l(\mathbf{x}) \sim \mathcal{GP}(\nu_l(\mathbf{x}), \gamma_l(\mathbf{x}, \mathbf{x}')), \quad l = 1, \dots, L, \quad (5.1)$$

where $\nu_l(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}$ and $\gamma_l(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ are a mean function and a covariance function, respectively, for the l -th latent GP $g_l(\mathbf{x})$, both of which are assumed integrable. Defining $f_s(\mathbf{x})$ as the s -th GP, the S -dimensional dependent GP $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_S(\mathbf{x}))^\top$ is assumed to be modeled as a linear mixing of the L independent latent GPs, then $\mathbf{f}(\mathbf{x})$ is given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}\mathbf{g}(\mathbf{x}) + \mathbf{n}(\mathbf{x}), \quad (5.2)$$

where $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_L(\mathbf{x}))^\top$, and where $\mathbf{W}$ is an $S \times L$ weight matrix in which (s, l) -entry $w_{s,l} \in \mathbb{R}$ is the weight of the l -th latent GP in the s -th data set, and where

$$\mathbf{n}(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, \mathbf{\Lambda}(\mathbf{x}, \mathbf{x}')) \quad (5.3)$$

is an S -dimensional zero-mean Gaussian noise process. Here, $\mathbf{0}$ is a column vector of 0's and $\mathbf{\Lambda}(\mathbf{x}, \mathbf{x}') = \text{diag}(\lambda_1(\mathbf{x}, \mathbf{x}'), \dots, \lambda_S(\mathbf{x}, \mathbf{x}'))$ with $\lambda_s(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ being a covariance function for the s -th Gaussian noise process. It is known that a linear combination of multivariate GPs is again a multivariate GP; thus, $f(\mathbf{x})$ can be simply written by

$$f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{m}(\mathbf{x}), \mathbf{K}(\mathbf{x}, \mathbf{x}')), \quad (5.4)$$

where the mean function $\mathbf{m}(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}^S$ is given by

$$\mathbf{m}(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] = \mathbf{W}\mathbb{E}[g(\mathbf{x})] = \mathbf{W}\boldsymbol{\nu}(\mathbf{x}), \quad (5.5)$$

and where the covariance function $\mathbf{K}(\mathbf{x}, \mathbf{x}') : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}^{S \times S}$ is given by

$$\begin{aligned} \mathbf{K}(\mathbf{x}, \mathbf{x}') &= \mathbb{E} \left[(f(\mathbf{x}) - \mathbf{W}\boldsymbol{\nu}(\mathbf{x}))(f(\mathbf{x}') - \mathbf{W}\boldsymbol{\nu}(\mathbf{x}'))^\top \right] \\ &= \mathbb{E} \left[(\mathbf{W}(g(\mathbf{x}) - \boldsymbol{\nu}(\mathbf{x})) + \mathbf{n}(\mathbf{x}))(\mathbf{W}(g(\mathbf{x}') - \boldsymbol{\nu}(\mathbf{x}')) + \mathbf{n}(\mathbf{x}'))^\top \right] \\ &= \mathbf{W}\mathbf{\Gamma}(\mathbf{x}, \mathbf{x}')\mathbf{W}^\top + \mathbf{\Lambda}(\mathbf{x}, \mathbf{x}'). \end{aligned} \quad (5.6)$$

Here, $\boldsymbol{\nu}(\mathbf{x}) = (\nu_1(\mathbf{x}), \dots, \nu_L(\mathbf{x}))^\top$ and $\mathbf{\Gamma}(\mathbf{x}, \mathbf{x}') = \text{diag}(\gamma_1(\mathbf{x}, \mathbf{x}'), \dots, \gamma_L(\mathbf{x}, \mathbf{x}'))$. The (s, s') -entry of $\mathbf{K}(\mathbf{x}, \mathbf{x}')$ is given by

$$k_{s,s'}(\mathbf{x}, \mathbf{x}') = \delta_{s,s'}\lambda_s(\mathbf{x}, \mathbf{x}') + \sum_{l=1}^L w_{s,l}w_{s',l}\gamma_l(\mathbf{x}, \mathbf{x}'). \quad (5.7)$$

One observes that the covariance function (5.7) for the multivariate GP $f(\mathbf{x})$ is represented by a linear combination of the covariance functions $\{\gamma_l(\mathbf{x}, \mathbf{x}')\}_{l=1}^L$ for the latent GPs, namely the covariance functions for latent GPs are shared among all data sets. Accordingly, different from the case of 2-stage SAGP model (See equation (4.15) in Chapter 4), the multivariate modeling allows us to effectively learn the spatial correlation for each data set by considering the dependences between data sets; this is more helpful in such a case that the numbers of observations in some data sets are small. In this thesis we focus on the case $L < S$, with the aim of reducing the number of free parameters as this helps to avoid overfitting (Teh, Seeger, and Jordan, 2005).

To handle the multivariate aggregated data, we introduce an observation model with a spatial aggregation process for each of the data sets. Multivariate aggregated observation is assumed to be obtained by integrating the mixed GP $f(\mathbf{x})$ over the corresponding

TABLE 5.1: Notation for SAGP-M.

Symbol	Description
L	number of latent GPs
l	index of latent GP, $l = 1, \dots, L$
$g_l(\mathbf{x})$	the l -th latent GP
$\nu_l(\mathbf{x})$	mean function of $g_l(\mathbf{x})$
$\gamma_l(\mathbf{x}, \mathbf{x}')$	covariance function of $g_l(\mathbf{x})$
$\mathbf{f}(\mathbf{x})$	S -dimensional dependent GP
$\mathbf{W}$	weight matrix of the latent GPs in the respective data sets, $\mathbf{W} \in \mathbb{R}^{S \times L}$
$\mathbf{m}(\mathbf{x})$	mean function of $\mathbf{f}(\mathbf{x})$
$\mathbf{K}(\mathbf{x}, \mathbf{x}')$	covariance function of $\mathbf{f}(\mathbf{x})$
$\mathbf{n}(\mathbf{x})$	S -dimensional zero-mean Gaussian noise process
$\lambda_s(\mathbf{x}, \mathbf{x}')$	covariance function for the s -th Gaussian noise process
σ_s^2	noise variance for the s -th data set

region; $\mathbf{y}$ is generated from the Gaussian distribution ¹,

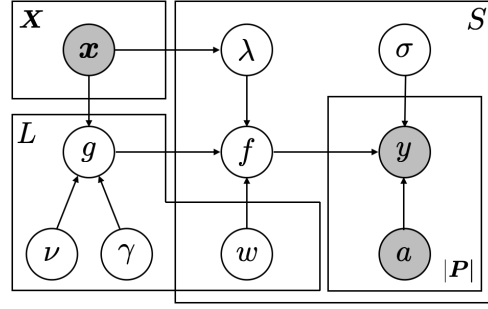
$$\mathbf{y} \mid \mathbf{f}(\mathbf{x}) \sim \mathcal{N} \left(\mathbf{y} \mid \int_{\mathbf{X}} \mathbf{A}(\mathbf{x}) \mathbf{f}(\mathbf{x}) d\mathbf{x}, \mathbf{\Sigma} \right), \quad (5.8)$$

where $\mathbf{A}(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}^{N \times S}$ is represented by

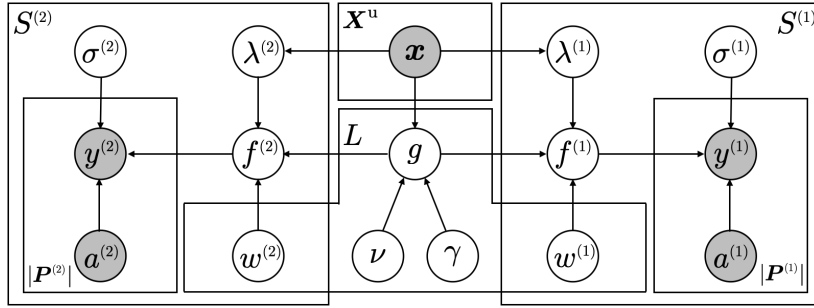
$$\mathbf{A}(\mathbf{x}) = \begin{pmatrix} \mathbf{a}_1(\mathbf{x}) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{a}_2(\mathbf{x}) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{a}_S(\mathbf{x}) \end{pmatrix}, \quad (5.9)$$

in which $\mathbf{a}_s(\mathbf{x}) : \mathbf{X} \rightarrow \mathbb{R}^{|P_s|}$ is a nonnegative weight function for spatial aggregation of the s -th data set. Here, $\mathbf{\Sigma} = \text{diag}(\sigma_1^2 \mathbf{I}, \dots, \sigma_S^2 \mathbf{I})$ in (5.8) is an $N \times N$ block diagonal matrix, where σ_s^2 is the noise variance for the s -th GP. Figure 5.2(a) shows a graphical model representation of SAGP-M in the case of a single domain, where shaded and unshaded nodes indicate observed and latent variables, respectively.

¹We assume that the integral appearing in (5.8) is well-defined. The assumption of the integrability of $\nu_l(\mathbf{x}), l = 1, \dots, L$, assures integrability of the mean function $\mathbf{m}(\mathbf{x})$ (5.5). This allows one to argue the conditions for the observation model (5.8) to be well-defined in a way similar to that in Appendix A.



(a) The case of a single domain.



(b) The case of two domains.

FIGURE 5.2: Graphical model representation of SAGP-M.

Extension to the case of multiple domains

One merit of SAGP-M is that it can be used for analyzing aggregated data sets across multiple domains (e.g., cities), by assuming that observations in each domain are conditionally independent given the latent GPs $\{g_l(\mathbf{x})\}_{l=1}^L$.

Let $v = 1, \dots, V$ denote a domain index, and $f^{(v)}(\mathbf{x})$ denote the multivariate GP for the v -th domain. The multivariate GPs for V domains are given by

$$f^{(v)}(\mathbf{x}) \sim \mathcal{GP}(\mathbf{m}^{(v)}(\mathbf{x}), \mathbf{K}^{(v)}(\mathbf{x}, \mathbf{x}')), \quad v = 1, \dots, V, \quad (5.10)$$

where the mean function $\mathbf{m}^{(v)}(\mathbf{x})$ and the covariance function $\mathbf{K}^{(v)}(\mathbf{x}, \mathbf{x}')$ are given by

$$\mathbf{m}^{(v)}(\mathbf{x}) = \mathbf{W}^{(v)} \mathbf{v}(\mathbf{x}), \quad (5.11)$$

$$\mathbf{K}^{(v)}(\mathbf{x}, \mathbf{x}') = \mathbf{W}^{(v)} \mathbf{\Gamma}(\mathbf{x}, \mathbf{x}') \mathbf{W}^{(v)\top} + \mathbf{\Lambda}^{(v)}(\mathbf{x}, \mathbf{x}'), \quad (5.12)$$

respectively. Here, $\mathbf{W}^{(v)}$ is the weight matrix for the v -th domain, and $\mathbf{\Lambda}^{(v)}(\mathbf{x}, \mathbf{x}')$ is the covariance function for the Gaussian noise process of the v -th domain.

Let $\mathbf{y}^{(v)}$ denote the observations for the v -th domain. The generative process of $\mathbf{y}^{(v)}$ is given by

$$\mathbf{y}^{(v)} \mid \mathbf{f}^{(v)}(\mathbf{x}) \sim \mathcal{N} \left(\mathbf{y}^{(v)} \mid \int_{\mathbf{X}^{(v)}} \mathbf{A}^{(v)}(\mathbf{x}) \mathbf{f}^{(v)}(\mathbf{x}) d\mathbf{x}, \boldsymbol{\Sigma}^{(v)} \right), \quad (5.13)$$

where $\mathbf{A}^{(v)}(\mathbf{x})$ is the weight function for spatial aggregation of the data sets in the v -th domain, and where $\boldsymbol{\Sigma}^{(v)}$ is the noise variances for the v -th domain. $\mathbf{X}^{(v)}$ is the input space for the v -th domain.

In the above formulation, the V data sets $\{\mathbf{y}^{(v)}\}_{v=1}^V$ are not directly correlated across domains; however, the shared covariance functions $\{\gamma_l(\mathbf{x}, \mathbf{x}')\}_{l=1}^L$ for the latent GPs can be learned by transfer learning based on the data sets from multiple domains. Thus, the spatial correlation for each data set could be more appropriately estimated via the shared covariance functions, which is more effective, especially in the case where we have only a few data sets available for a single domain. The graphical model representation of SAGP-M for the case of two domains is shown in Figure 5.2(b), where $\mathbf{X}^u$ is the union of the input spaces for both domains.

5.2.2 Inference

Given the multiple aggregated data sets, the posterior GP with multiple outputs is derived on the basis of a Bayesian inference procedure. In the following, we briefly describe the marginal likelihood for the cases of a single domain and multiple domains, respectively; the posterior GP is then constructed. Derivations of the marginal likelihood and the posterior GP can be obtained in a manner similar to those of SAGP-S described in Chapter 3. See also Supplementary Material of (Tanaka et al., 2019b).

Marginal likelihood for a single domain

The model parameters $\mathbf{W}$, $\boldsymbol{\Lambda}(\mathbf{x}, \mathbf{x}')$, $\boldsymbol{\Sigma}$, $\nu(\mathbf{x})$, $\boldsymbol{\Gamma}(\mathbf{x}, \mathbf{x}')$ are estimated on the basis of the marginal likelihood, which we describe below. Given the aggregated data sets $\mathbf{D}$, the marginal likelihood of $\mathbf{y}$ is given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{C}), \quad (5.14)$$

where we analytically integrate out the GP prior $f(\mathbf{x})$. $\boldsymbol{\mu}$ is an N -dimensional mean vector represented by

$$\boldsymbol{\mu} = \int_{\mathbf{X}} \mathbf{A}(\mathbf{x}) \mathbf{m}(\mathbf{x}) d\mathbf{x}, \quad (5.15)$$

which is the integral of the mean function $\mathbf{m}(\mathbf{x})$ of $f(\mathbf{x})$ over the respective regions for all data sets. $\mathbf{C}$ is an $N \times N$ covariance matrix represented by

$$\mathbf{C} = \iint_{\mathbf{X} \times \mathbf{X}} \mathbf{A}(\mathbf{x}) \mathbf{K}(\mathbf{x}, \mathbf{x}') \mathbf{A}(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \boldsymbol{\Sigma}. \quad (5.16)$$

It is an $S \times S$ block matrix whose (s, s') -th block $\mathbf{C}_{s,s'}$ is a $|\mathbf{P}_s| \times |\mathbf{P}_{s'}|$ matrix represented by

$$\mathbf{C}_{s,s'} = \iint_{\mathbf{X} \times \mathbf{X}} k_{s,s'}(\mathbf{x}, \mathbf{x}') \mathbf{a}_s(\mathbf{x}) \mathbf{a}_{s'}(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \delta_{s,s'} \sigma_s^2 \mathbf{I}. \quad (5.17)$$

The diagonal blocks (i.e., $s = s'$) provide the region-to-region covariance matrices in the respective data sets, each of which is equivalent to (3.6) in the single output case. In SAGP-M, one can naturally define the off-diagonal blocks (i.e., $s \neq s'$), that is, the region-to-region covariance matrices across data sets, which allow us to incorporate the dependences between data sets considering the spatial correlations even if the observed data sets are associated with the partitions of different spatial granularities.

Marginal likelihood for multiple domains

In SAGP-M, the observations for different domains are assumed to be conditionally independent given the shared latent GPs $\{g_l(\mathbf{x})\}_{l=1}^L$. The marginal likelihood of $\{\mathbf{y}^{(v)}\}_{v=1}^V$ is then given by the product of those for the V domains, represented by

$$p(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(V)}) = \prod_{v=1}^V \mathcal{N}(\mathbf{y}^{(v)} | \boldsymbol{\mu}^{(v)}, \mathbf{C}^{(v)}), \quad (5.18)$$

where the mean vector $\boldsymbol{\mu}^{(v)}$ and the covariance matrix $\mathbf{C}^{(v)}$ for the v -th domain are given by

$$\boldsymbol{\mu}^{(v)} = \int_{\mathbf{X}^{(v)}} \mathbf{A}^{(v)}(\mathbf{x}) \mathbf{m}^{(v)}(\mathbf{x}) d\mathbf{x}, \quad (5.19)$$

$$\mathbf{C}^{(v)} = \iint_{\mathbf{X}^{(v)} \times \mathbf{X}^{(v)}} \mathbf{A}^{(v)}(\mathbf{x}) \mathbf{K}^{(v)}(\mathbf{x}, \mathbf{x}') \mathbf{A}^{(v)}(\mathbf{x}')^\top d\mathbf{x} d\mathbf{x}' + \boldsymbol{\Sigma}^{(v)}, \quad (5.20)$$

respectively. By maximizing the logarithm of (5.18), one can estimate the model parameters utilizing all data sets from multiple domains via the shared covariance function $\Gamma(\mathbf{x}, \mathbf{x}')$ contained in $\mathbf{K}^{(v)}(\mathbf{x}, \mathbf{x}')$ (5.12).

Posterior GP

We have only to consider the case of a single domain, because the derivation of the posterior GP can be conducted independently for each domain. Given the aggregated data sets D and the estimated parameters, the posterior GP $f^*(x)$ with S outputs is given by

$$f^*(x) \sim \mathcal{GP}(m^*(x), K^*(x, x')), \quad (5.21)$$

where $m^*(x) : X \rightarrow \mathbb{R}^S$ and $K^*(x, x') : X \times X \rightarrow \mathbb{R}^{S \times S}$ are the mean function and the covariance function for $f^*(x)$, respectively. The mean function $m^*(x)$ and the covariance function $K^*(x, x')$ are given by

$$m^*(x) = m(x) + H(x)^\top C^{-1}(y - \mu), \quad (5.22)$$

$$K^*(x, x') = K(x, x') - H(x)^\top C^{-1}H(x'), \quad (5.23)$$

respectively, where

$$H(x) = \int_X A(x')K(x', x) dx', \quad (5.24)$$

which consists of the point-to-region covariances, that is, the covariances between any location point x and the respective regions in all data sets. Once the posterior GP (5.21) is obtained, one can predict multivariate data values at any location point; and also output the prediction and its variance at the region by integrating the posterior GP over the corresponding region.

5.3 Experiments

5.3.1 Experimental setup

The experimental setting is the same as that described in Section 4.3. We evaluated the proposed model using multiple real-world data sets from two cities: New York City and Chicago. Details of the data sets are described in Section 4.3.1. We examined the refinement task (as described in Section 4.3.2) to show the effectiveness of the proposed model. The prediction performance in the data refinement was evaluated using the metric: mean absolute percentage error (MAPE) of the fine-grained target data.

5.3.2 Setup of SAGP-M

In the experiments, we used zero-mean Gaussian processes as the latent GPs $\{g_l(\mathbf{x})\}_{l=1}^L$, i.e., $v_l(\mathbf{x}) = 0$ for $l = 1, \dots, L$. We used the following squared-exponential kernel as the covariance function for the latent GPs $\{g_l(\mathbf{x})\}_{l=1}^L$ and for the Gaussian noise process $\mathbf{n}(\mathbf{x})$,

$$\gamma_l(\mathbf{x}, \mathbf{x}') = \alpha_l^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\beta_l^2}\right), \quad l = 1, \dots, L, \quad (5.25)$$

$$\lambda_s(\mathbf{x}, \mathbf{x}') = \alpha_s^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\beta_s^2}\right), \quad s = 1, \dots, S. \quad (5.26)$$

α_l^2 and α_s^2 are signal variances for the l -th latent GP and for the s -th Gaussian noise process, respectively, where we set $\alpha_l^2 = 1$ because the variance can already be represented by scaling the columns of $\mathbf{W}$. β_l and β_s are scale parameters for the l -th latent process and for the s -th noise process, respectively.

The model parameters, $\mathbf{W}$, $\{\beta_l\}$, $\{\alpha_s\}$, $\{\beta_s\}$, Σ , were learned by maximizing the logarithm of the marginal likelihood (5.14) or (5.18) using the L-BFGS method (Liu and Nocedal, 1989). In order to approximate the integral over regions, we used sufficiently fine-grained square grid cells, the size of which was $300 \text{ m} \times 300 \text{ m}$ for both cities. The number L of the latent GPs was determined from $\{1, \dots, S - 1\}$ via leave-one-out cross-validation (Bishop, 2006); the validation error was obtained using each held-out coarse-grained data value.

5.3.3 Baselines

The proposed model, SAGP-M, was compared with SAGP-S, 2-stage SAGP model, and semiparametric latent factor model (SLFM) (Teh, Seeger, and Jordan, 2005). The setup of SAGP-S and 2-stage SAGP model was the same as that described in Section 4.3.4. SLFM is an instance of the linear model of coregionalization (LMC) (Álvarez, Rosasco, and Lawrence, 2012), in which multiple dependent outputs are defined as a linear mixing of latent GPs. SLFM assumes that each sample is observed at a location point; we then regarded each aggregated observation as being associated with the centroid of the corresponding region. SAGP-M is the extension of SLFM; it can naturally handle the aggregated observations by incorporating the observation model with spatially aggregation processes.

5.3.4 Results for the case of a single city

Refinement performance

Table 5.2 shows MAPE and standard errors for SAGP-S, 2-stage SAGP model, SLFM, and SAGP-M. For all data sets, SAGP-M yielded the comparable or better performance than the other models; the differences between SAGP-M and the baselines were statistically tested by Student's t-test. The MAPEs of SAGP-M were lower than those of the regression-based model (i.e., 2-stage SAGP model) for almost all data sets, showing that multivariate approach is a more effective way for modeling multiple aggregated data sets. For all data sets, SAGP-M yielded lower MAPEs compared with the baseline model, i.e., SLFM; these results indicate that the key technique, the observation model with spatial aggregation processes, is essential in refining coarse-grained aggregated data accurately.

Figures 5.3 and 5.4 show the refinement results for poverty rate data and recycle diversion rate data, respectively, in New York City; the true data are shown on the left, and the predictions of SAGP-M, SLFM, 2-stage SAGP, and SAGP-S are shown on the right. Here, we normalized the values of each model to the range $[0, 1]$. As shown in these figures, SAGP-M yielded better refinements compared with the other models.

TABLE 5.2: Performance comparisons for SAGP-M: MAPE and standard errors for the prediction of fine-grained data in New York City and Chicago. The numbers in parentheses denote the number L of the latent GPs estimated by the validation procedure. The single star (*) and the double star (**) indicate significant difference between SAGP-M and the other models at the levels of P values of < 0.05 and < 0.01 , respectively.

(a) New York City				
	SAGP-S	2-stage SAGP model	SLFM	SAGP-M
PM2.5	0.053 ± 0.007	0.042 ± 0.005	0.036 ± 0.005 (6)	$0.030 \pm 0.005^*$ (5)
Poverty rate	0.281 ± 0.034	0.218 ± 0.021	0.207 ± 0.025 (4)	$0.181 \pm 0.021^{**}$ (3)
Unemployment rate	0.232 ± 0.025	0.154 ± 0.021	0.195 ± 0.024 (3)	0.165 ± 0.020 (4)
Mean commute	0.076 ± 0.010	0.052 ± 0.005	0.057 ± 0.007 (4)	0.048 ± 0.007 (5)
Population	0.371 ± 0.040	0.321 ± 0.039	0.337 ± 0.039 (3)	$0.282 \pm 0.034^*$ (4)
Recycle diversion rate	0.282 ± 0.034	0.191 ± 0.023	0.222 ± 0.032 (4)	$0.162 \pm 0.022^{**}$ (4)
Crime	0.525 ± 0.112	0.413 ± 0.086	0.401 ± 0.053 (2)	$0.334 \pm 0.048^{**}$ (2)
Fire incident	0.793 ± 0.148	0.560 ± 0.106	0.500 ± 0.052 (4)	$0.402 \pm 0.040^*$ (5)
311 call	0.069 ± 0.005	0.061 ± 0.004	0.061 ± 0.004 (6)	$0.050 \pm 0.004^{**}$ (3)
Public telephone	0.099 ± 0.008	0.089 ± 0.008	0.086 ± 0.008 (4)	0.079 ± 0.008 (5)
(b) Chicago				
	SAGP-S	2-stage SAGP model	SLFM	SAGP-M
Poverty rate	0.423 ± 0.063	0.274 ± 0.035	0.335 ± 0.052 (2)	$0.255 \pm 0.030^{**}$ (2)
Unemployment rate	0.405 ± 0.041	0.257 ± 0.030	0.278 ± 0.025 (2)	$0.228 \pm 0.021^*$ (2)

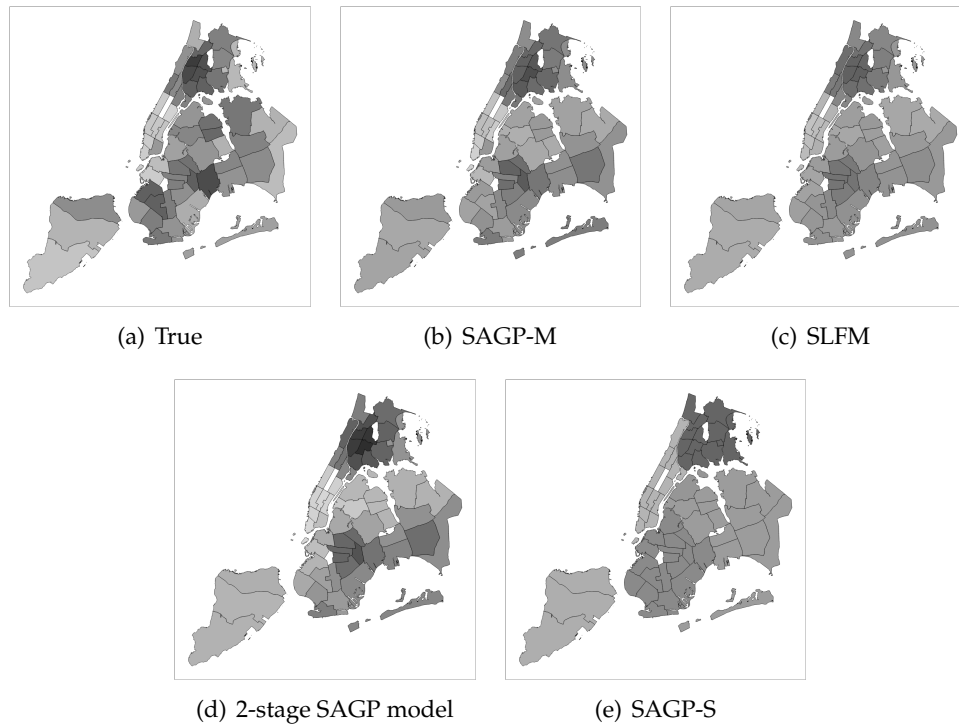


FIGURE 5.3: Comparison of the refined poverty rate data in New York City.

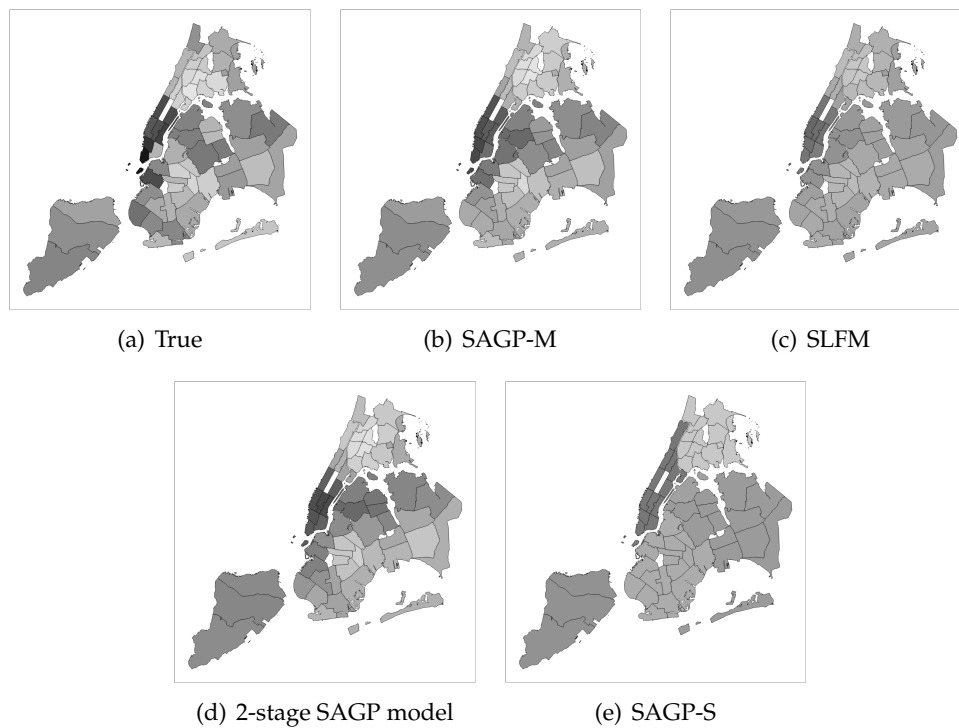


FIGURE 5.4: Comparison of the refined recycle diversion rate data in New York City.

TABLE 5.3: Comparisons of computational time (second).

(a) New York City			
	2-stage SAGP model	SLFM	SAGP-M
PM2.5	211.3	1986.6	3693.4
Poverty rate	234.5	2139.0	2692.2
Unemployment rate	228.5	2151.4	3618.4
Mean commute	200.2	1732.4	3016.9
Population	213.4	1606.1	3246.6
Recycle diversion rate	234.6	1256.0	2973.7
Crime	208.4	1762.2	2949.4
Fire incident	184.7	1850.8	2827.5
311 call	179.0	1793.0	3055.6
Public telephone	172.1	1220.8	2071.7

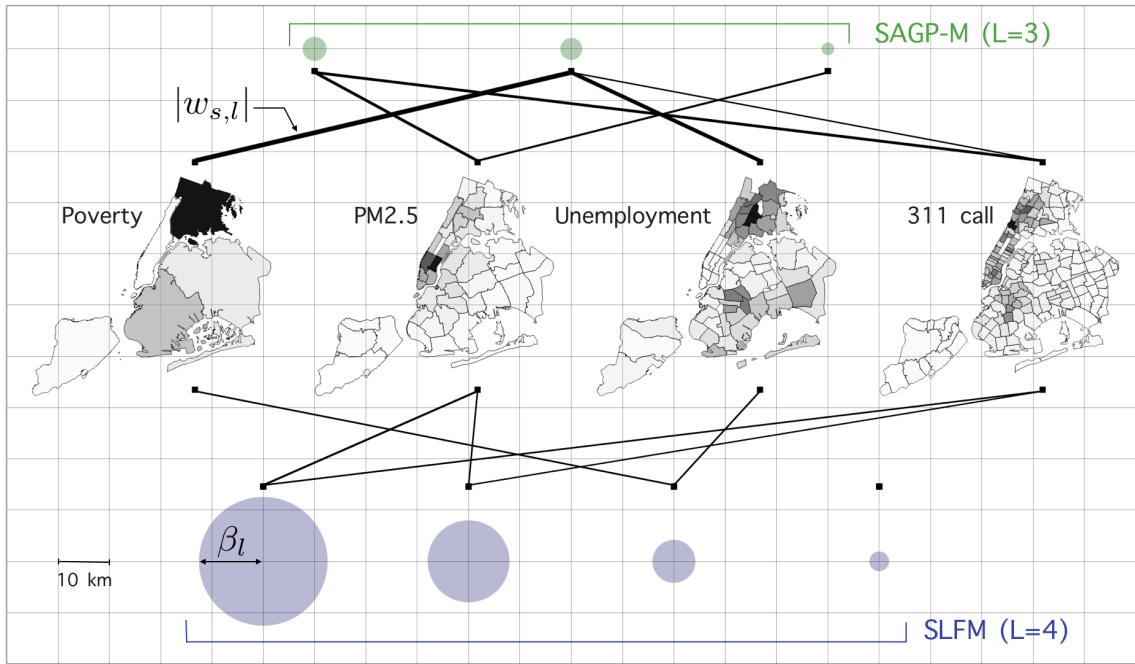
(b) Chicago			
	2-stage SAGP model	SLFM	SAGP-M
Poverty rate	21.7	31.7	364.3
Unemployment rate	23.1	32.6	326.8

Computation time

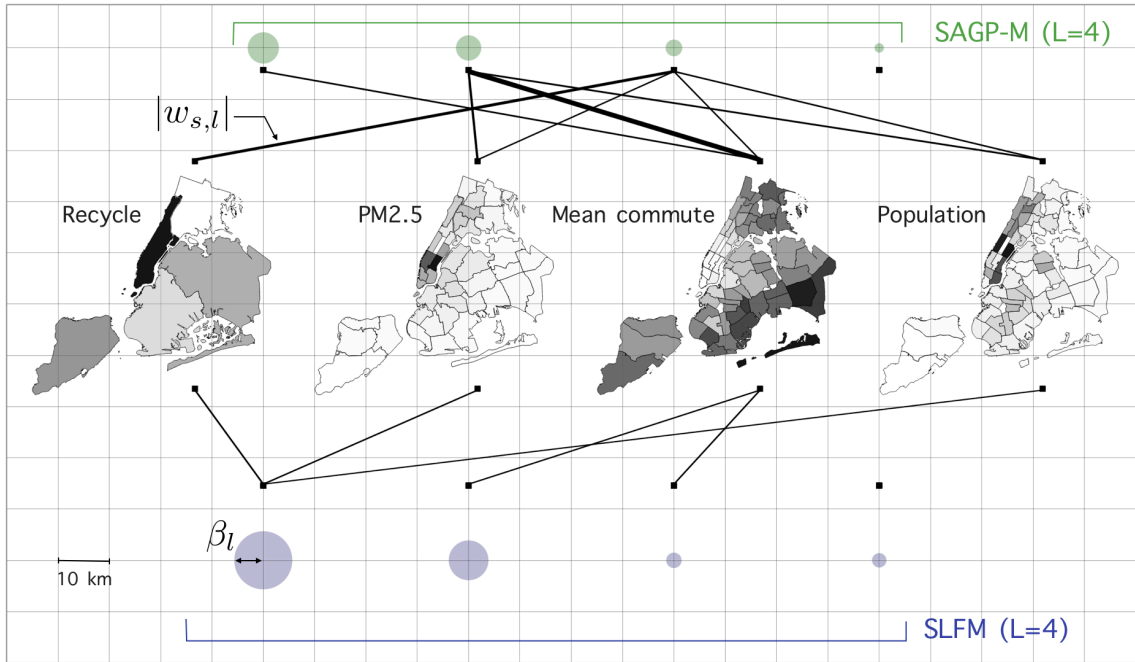
Table 5.3 shows the computational time of the inference process for 2-stage SAGP model, SLFM, and SAGP-M, where the experiments were conducted on a 3.1 GHz Intel Core i7. As shown in this table, 2-stage SAGP model has an advantage in the computational efficiency, which is reasonable because the inference of 2-stage SAGP model can be independently performed for each data set (as described in Section 4.2.2). One observes that the computational time of SAGP-M drastically increased than that of 2-stage SAGP model; instead, SAGP-M can improve the prediction performance as shown in Table 5.2.

Parameter visualization

Figures 5.5(a) and 5.5(b) visualize the mixing weights $\mathbf{W}$ and the scale parameters $\{\beta_l\}_{l=1}^L$ estimated by SAGP-M and SLFM in the cases of refining poverty rate and recycle diversion rate data sets, respectively, in New York City. In Figures 5.5(a) and 5.5(b), we picked up 4 aggregated data sets in the case of refining each target data set, and visualize their input data. In both figures, one observes that the scale parameters estimated by SAGP-M are relatively small compared with those estimated by SLFM, presumably because the spatial



(a) The case of refining poverty rate data in New York City.



(b) The case of refining recycle diversion rate data in New York City.

FIGURE 5.5: Visualization of the estimated parameters. The radii of green and blue circles equal the values of β_l estimated by SAGP-M and SLFM, respectively. The edge widths are proportional to the absolute weights $|w_{s,l}|$ estimated by the respective models. Here, we omitted those edges whose absolute weights were lower than a threshold.

TABLE 5.4: Performance comparisons for transfer learning across two cities.

	Chicago	
	Poverty rate	Unemployment rate
SLFM	0.335 ± 0.052 (2)	0.278 ± 0.025 (2)
SLFM (trans)	0.328 ± 0.050 (6)	0.255 ± 0.024 (5)
SAGP-M	0.255 ± 0.030 (2)	0.228 ± 0.021 (2)
SAGP-M (trans)	0.223 ± 0.022 (4)	0.210 ± 0.019 (5)

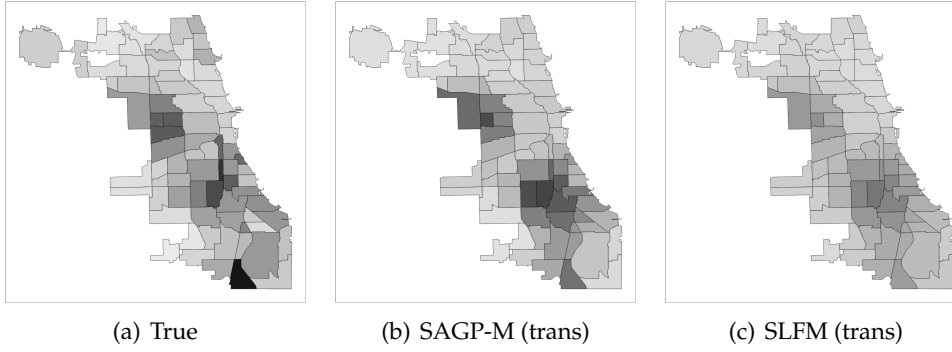


FIGURE 5.6: Comparison of the refined poverty rate data in Chicago (transfer learning across two cities).

aggregation process incorporated in SAGP-M effectively separates intrinsic spatial correlations and apparent smoothing effects due to the spatial aggregation to yield aggregated observations. The comparisons of the estimated weights show that SAGP-M emphasized the useful dependences between data sets, e.g., the strong correlation between poverty rate data and unemployment rate data (See Figure 5.5(a)).

5.3.5 Results for the case of two cities

In our formulation, multivariate dependent GP is constructed by linearly mixing independent latent GPs, one benefit of which is that it can be straightforwardly used for transfer learning across multiple domains (i.e., cities). This functionality is more effective in such a case where we have only a few data sets available on a single city. We here show the results of refining poverty rate and unemployment rate data sets in Chicago with simultaneously utilizing the data sets from New York City.

Table 5.4 shows MAPE and standard errors for SLFM, SLFM (trans), SAGP-M, and SAGP-M (trans). Figure 5.6 illustrates the prediction results for poverty rate data in Chicago made by each model. One observes that SAGP-M (trans) provided the lowest MAPE in the compared methods. These results indicate that our proposed model can improve the

refinement performance by sharing knowledge across data sets from different cities. Our model can still be applicable even if there are only a few data sets available on the target city.

5.4 Discussion

In this chapter, we proposed a novel multivariate Gaussian process (GP) model, spatially aggregated Gaussian process with multiple outputs (SAGP-M), that can utilize multiple aggregated data sets with various granularities to obtain the functions for respective data sets. In SAGP-M, we assume that a dependent multivariate GP is expressed as a linear mixing of several independent latent GPs. One can derive the posterior GP with multivariate outputs given multiple aggregated data sets with various granularities via the introduction of an observation model for integral observations at respective regions.

In the experiments, we demonstrated that SAGP-M can 1) accurately refine coarse-grained aggregated data than the compared models and 2) provide performance improvements by utilizing data sets from multiple cities.

There are several important research issues that should be addressed in the future. First, the computational speed-up is required for large-scale data sets. One promising solution is to adopt a scalable variational inference with inducing points, as in (Titsias, 2009). Second, it would be interesting to consider a non-stationary model, which could be achieved by considering an input-dependent weight function $w_{s,l}(x)$. The model proposed in (Hamelijnck et al., 2019) has defined the weight function as a GP. Such formulations could be used for representing a non-stationary GP as a mixing of independent and stationary latent GPs by weight functions. Third, the proposed formulation offers a general framework for modeling aggregated data, thus it may be applied to various applications. For example, our model can be straightforwardly used for analyzing time-series data (e.g., sensor data) that have been aggregated over time-intervals, as in (Yousefi, Smith, and Álvarez, 2019).

Although the encouraging results for the case of two cities have already been obtained in our experiments, we believe that this is the first step toward establishing a transfer learning framework from aggregated data sets across multiple domains. In the present model, only latent GPs are shared among domains; one of our future works is to introduce a mechanism of inferring the relationships between domains, which could be helpful in estimating model parameters (e.g., mixing weights). We should also conduct experiments using aggregated data sets from more domains.

Part II

Inferring Latent People Flow from Aggregated Population Data

Chapter 6

Inferring Latent People Flow from Aggregated Data at Limited Locations

6.1 Introduction

Recent advances in wireless and mobile networks have generated an explosion of location information of people in various spaces including exhibition halls, amusement parks, shopping malls, and urban cities. In this chapter, we focus on people's mobility behaviors in a space of interest. It is very important to uncover people's mobility patterns from location data in many applications such as navigation systems (Huang and Gartner, 2010), travel route recommendation (Kurashima et al., 2010), location-based marketing (Dhar and Varshney, 2011), urban planning (Yuan, Zheng, and Xie, 2012), and disaster management (Song et al., 2014). For instance, finding popular routes of people in exhibition halls yields better route recommendations. Discovering transition patterns in urban areas is helpful for optimizing public transportation and communication systems.

There have been many statistical methods that analyze individual-level data, i.e., *trajectory data*, that is a collection of individual location and time (Giannotti et al., 2007; Monreale et al., 2009; Zhuang et al., 2017). However, the individual-level data are often aggregated over regions (e.g., sensor ranges and grid cells) for preserving people's privacy; then aggregated population data are only available. The use of aggregated population data is also advantageous in that the data can be obtained directly by people counting systems that use various sensors such as video cameras (Chan, Liang, and Vasconcelos, 2008) and inductive-loop traffic detectors (Klein, 2001); these are low-cost and easy to install. Giving the preference for data aggregation, existing methods that analyze trajectory data are inapplicable.

In this study, we assume that the aggregated population data consist of the numbers of incoming and outgoing people at each observed location and at each time step. In this

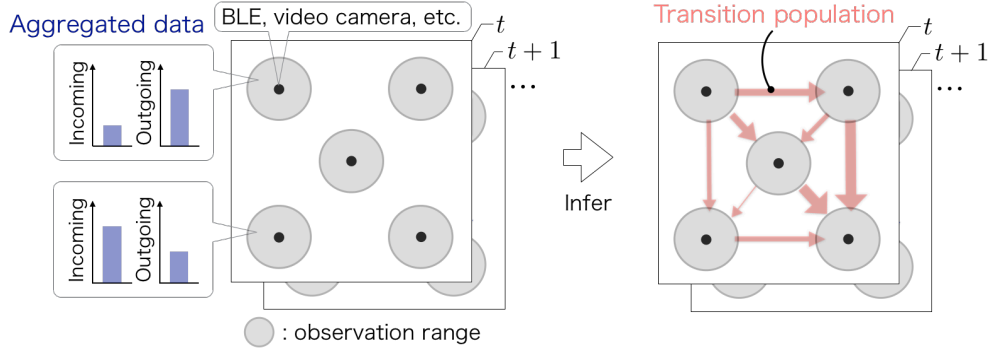


FIGURE 6.1: The problem we focus on: Inferring latent people flows between locations from aggregated population data. Five black dots represent observed locations; gray circles are observation ranges (or predefined regions), in which people can be counted. (Left figure) The numbers of incoming and outgoing people at each location and at each time step, which we use as input data. (Right figure) The output to be estimated: unobserved transition populations between locations.

chapter, we simply call the regions as *locations*, where it should be noted that it does not mean location points at the individual-level. These data do not have tracking information of individuals. The left part of Figure 6.1 shows an example of the data collected in a space of interest. For example, in an exhibition hall, we may be able to obtain the numbers of visitors entering or leaving each event booth at each time step via wireless technologies, e.g., Bluetooth Low Energy (BLE). Other examples include pedestrian data for each attraction in an amusement park (Du, Kumar, and Varakantham, 2014), each store in a shopping mall (Senior et al., 2007), and traffic data from intersections (Kumar, Sheldon, and Srivastava, 2013). We consider the problem of estimating latent people flows, that is, transition populations between observed locations, given just aggregated population data (See the right part of Figure 6.1). Solving this problem allows us to utilize aggregated population data for understanding people’s mobility patterns in a space of interest.

This problem has recently been addressed in the framework of *collective flow diffusion models* (CFDM) (Kumar, Sheldon, and Srivastava, 2013; Du, Kumar, and Varakantham, 2014; Iwata et al., 2017; Akagi et al., 2018; Iwata and Shimizu, 2019). In CFDM, people flows are assumed to be probabilistic diffusion processes on a location graph. Transition populations are estimated by maximizing a likelihood subject to constraints that represent flow conservation law. The constraints state that any person who leaves a location at one time step always arrives at another location at the same time step. In other words, existing models assume that everyone is always in one of the observed locations at every time step. However, this assumption is too restrictive for many real-world settings, because we do not always have a large enough number of sensor devices to cover a large-scale space of interest: Each sensor device has an observation range in which people can be counted (depicted by the gray circles in Figure 6.1), meaning that people might be moving through passages that are outside the ranges at an observation time step. In that case, the flow

conservation constraints might not be satisfied, and thus CFDM fails to estimate transition populations between locations accurately.

We propose a probabilistic model, called *Time-delayed Collective Flow Diffusion Model (T-CFDM)*, that can estimate transition populations from the numbers of incoming and outgoing people gathered from limited locations. We newly introduce an observation model with aggregation processes that incorporate travel duration distributions between observed locations, which aims to model the people in transit, those who leave one location in one time step and arrive at another location after some delay. An important characteristic of this formulation is to treat travel duration as a random variable that follows a probability distribution; it does not use a point estimate. This enables us to capture the heterogeneity in travel duration among individuals; for instance, in an exhibition hall, some people might promptly move to their next location, but others might take a rest before arriving at another location; in urban areas, moving speeds might depend on the means of transport used (e.g., walking, bike, car).

We develop an approximate expectation-maximization algorithm for learning transition populations and model parameters, simultaneously. The estimation based on the observation model we introduce is robust in noisy settings, where the noise is regarded as a factor that may disturb flow conservation law. For example, people may enter or exit from the predefined space, implying that the total number of people is not constant.

We evaluate the effectiveness of T-CFDM by conducting extensive experiments on real-world data sets, pedestrian location logs from large-scale exhibition halls, and bike trip data and taxi trip data from New York City.

The remainder of this chapter is organized as follows. In Section 6.2, we summarize the related works. Section 6.3 describes our problem of inferring latent people flows from aggregated population data. In Section 6.4, we present the formulation of T-CFDM and its inference procedure. Section 6.5 shows the effectiveness of T-CFDM by using multiple real-world data sets. Finally, we present a summary of this chapter and a discussion of future work in Section 6.6.

6.2 Related work

6.2.1 Collective graphical models

Collective graphical models (CGMs) have recently been developed as a general framework for modeling aggregated data (Sheldon and Dietterich, 2011). They have been used for inferring contingency tables given just marginals. Contingency table estimation is also

known as the *ecological inference problem*, which has a long history and has attracted in the field of statistics (Robinson, 1950; King, 2013; Flaxman, Wang, and Smola, 2015). One benefit of CGMs is that one can use efficient probabilistic inference techniques based on the maximum a posteriori (Sheldon et al., 2013), MCMC sampler (Sheldon and Dietterich, 2011), message passing (Sun, Sheldon, and Kumar, 2015), and variational Bayesian inference (Iwata et al., 2017). Collective flow diffusion model (CFDM) is the first application of the inference techniques developed for CGMs to the transportation domain (Kumar, Sheldon, and Srivastava, 2013); and its variants have been developed for estimating transition populations between locations from just aggregated population data (Du, Kumar, and Varakantham, 2014; Iwata et al., 2017; Akagi et al., 2018; Iwata and Shimizu, 2019). One can obtain transition populations via a maximum likelihood estimation with the constraints that represent flow conservation law. However, these existing models require that everyone is always in one of the observed locations at every time step. To satisfy this requirement, a large number of sensor devices are needed to cover the large-scale space targeted; this is an unrealistic requirement.

Unlike the existing models, the proposed model incorporates travel duration distributions into the constraints that represent flow conservation law. This allows us to more accurately estimate the transition populations by considering the travel durations between observed locations even if we have just aggregated population data gathered from limited locations.

6.2.2 Information diffusion models

Information diffusion models are related to our proposal, because their main goal is to estimate latent flows of a piece of information over social networks (Kempe, Kleinberg, and Tardos, 2003; Leskovec, Adamic, and Huberman, 2007). Recently, several probabilistic models have been proposed for estimating latent networks and model parameters without prior knowledge on social network structures (Rodriguez, Balduzzi, and Schölkopf, 2011; Iwata, Shah, and Ghahramani, 2013; Kurashima et al., 2014; Tanaka et al., 2016). These models incorporate a continuous-time distribution for estimating information flows that spread from node to node with some delay. In the analysis of information flows, it is not, however, necessary to consider flow conservation law, a critical factor in modeling the diffusion processes of people flows.

Inspired by the idea of information diffusion modeling, we design travel duration distributions between locations. By maximizing the likelihood with flow conservation constraints that incorporate the travel duration distributions, our model can infer latent people flows between locations from aggregated population data.

6.2.3 Urban computing

There have been many statistical models for analyzing individual-level data, e.g., GPS trajectory data, collected by mobile devices (Giannotti et al., 2007; Monreale et al., 2009; Kurashima et al., 2010; Yuan, Zheng, and Xie, 2012). These models, however, require us to prepare the histories of locations visited by each user. Unfortunately, individual-level data are often aggregated to protect privacy, which makes it difficult to employ these models.

Many methods based on deep neural networks (DNNs) have recently been developed for analyzing aggregated population data (Hoang, Zheng, and Singh, 2016; Zhang, Zheng, and Qi, 2017; Yao et al., 2018; Yao et al., 2019). The problem addressed in these studies is related but essentially different from ours. The purpose of these models is, while incorporating various external factors (e.g., weather), to predict the future numbers of incoming and outgoing people at each location; they cannot estimate the transition populations between locations from just the numbers of incoming and outgoing people. A DNN-based model has been developed for estimating transition populations (Zhang et al., 2019); however, this model is learned in the supervised setting; namely one has to collect the transition populations between locations for training the model. Our study, on the other hand, considers the problem of estimating the transition populations from just the numbers of incoming and outgoing people, where the estimation process can be achieved in the unsupervised setting; namely it does not require the training data for transition populations. The unsupervised setting we focus on is practically important, because most of the aggregated population data gathered by cost-effective ways (e.g., people counting systems) do not have transition information.

The method of (Xu et al., 2017) tried to directly recover user trajectories given aggregated data and transition probabilities. The major drawback is that the transition probabilities must be manually set on the basis of additional auxiliary information such as distances between locations. The method cannot estimate the transition populations between locations from just aggregated population data.

Different from the prior works in the field of urban computing, our proposal achieves the transition population estimation given just the numbers of incoming and outgoing people. The estimation process can be conducted in the unsupervised setting, meaning that it does not require the training data for transition populations and the auxiliary information (e.g., distances). Moreover, our model can estimate travel durations between observed locations; this yields robust estimations even if the number of observed locations is limited.

6.3 Problem formulation

In this section, we describe the mathematical notations used in this chapter, and define our problem of inferring latent people flows from aggregated population data, i.e., the numbers of incoming and outgoing people at each location and at each time step. The notations are listed in Table 6.1.

Aggregated population data. Let $t \in \{1, \dots, T\}$ be a discrete time step. Let V denote a set of locations and $i \in V$ denote a location. In this chapter, we use the word *location* to represent an observed region; it does not mean a location point of an individual. Letting $\mathbb{N}^0 = \{0, 1, 2, \dots\}$ denote the set of nonnegative integers, let $y_{ti}^{\text{in}} \in \mathbb{N}^0$ and $y_{ti}^{\text{out}} \in \mathbb{N}^0$ be the numbers of incoming and outgoing people, respectively, of location i at time step t , where we assume that people are aggregated by each time interval $[t-1, t)$, namely people who left or entered a location during the interval $[t-1, t)$ are counted in the numbers at time step t . Aggregated population data refer to the paired sets of the outgoing people counts $Y^{\text{out}} = \{y_{ti}^{\text{out}} \mid t = 1, \dots, T; i \in V\}$ and of the incoming people counts $Y^{\text{in}} = \{y_{ti}^{\text{in}} \mid t = 1, \dots, T; i \in V\}$.

Problem. Our problem is to infer latent people flows, i.e., transition populations between locations, from aggregated population data; this problem is illustrated schematically in Figure 6.1. Let $G = (V, E)$ be the undirected graph that represents accessibility information in such a way that the neighbor $E_i = \{j \in V \mid (ij) \in E\}$ of location $i \in V$ on G represents the set of locations that are accessible from i . Let $z_{tij} \in \{0, \dots, y_{ti}^{\text{out}}\}$ be the transition population, that is, the number of people who left location $i \in V$ at time step t and whose next location is $j \in E_i$. Define the set of transition populations as $Z = \{z_{tij} \mid t = 1, \dots, T; i \in V; j \in E_i\}$. Given $Y^{\text{out}}, Y^{\text{in}}, T$, and G , we wish to infer the transition populations Z .

6.4 Proposed model

6.4.1 Model

We propose T-CFDM (Time-delayed Collective Flow Diffusion Models), a probabilistic model for inferring latent people flows, i.e., transition populations between locations, from aggregated population data gathered at limited locations. We first model the temporal dynamics of people flows, which are assumed to be probabilistic diffusion processes on a location graph. We then introduce an observation model that incorporates aggregation processes, which enables us to estimate transition populations while considering

TABLE 6.1: Notation for T-CFDM.

Symbol	Description
T	number of time steps
t	time step, $t \in \{1, \dots, T\}$
V	set of locations
G	undirected graph, $G = (V, E)$
i, j	locations, $i, j \in V$
E_i	set of locations accessible from location i , $E_i = \{j \in V \mid (ij) \in E\}$
$\mathbb{N}^0$	set of nonnegative integers, $\mathbb{N}^0 = \{0, 1, 2, \dots\}$
y_{ti}^{out}	number of outgoing people of location i at time step t , $y_{ti}^{\text{out}} \in \mathbb{N}^0$
y_{ti}^{in}	number of incoming people of location i at time step t , $y_{ti}^{\text{in}} \in \mathbb{N}^0$
m_{ti}^{out}	noise-free variable for y_{ti}^{out}
m_{ti}^{in}	noise-free variable for y_{ti}^{in}
z_{tij}	transition population leaving location i at time step t and whose next location is j , $z_{tij} \in \{0, \dots, y_{ti}^{\text{out}}\}$
θ_{ij}	transition probability that people move from location i to j , $\theta_{ij} \geq 0$, $\sum_{j \in E_i} \theta_{ij} = 1$
π_{ij}	set of parameters of the travel duration distribution from location i to j
σ_i^2	noise variance for the number of outgoing people at location i
η_i^2	noise variance for the number of incoming people at location i

flow conservation constraints. By incorporating travel duration distributions between locations into these constraints, our model can robustly infer transition populations in such a situation that the sensor devices have limited observation range and some people are not observed in any location in some time periods.

In T-CFDM, the temporal dynamics of people flows are assumed to be probabilistic diffusion processes on a graph, where the nodes are locations and the edges are paths between locations. Even if the graph structure such as a road network or a set of neighbor information is unavailable, our model is still applicable by assuming a complete graph among locations. We assume that each individual moves from location to location in accordance with time-independent transition probabilities that are shared among all individuals. Let $\theta_{ij} \geq 0$ be the transition probability that individuals move from location i to j , where $\sum_{j \in E_i} \theta_{ij} = 1$. Since the transition probabilities are shared among people, the transition populations $\mathbf{Z}_{ti} = \{z_{tij} \mid j = 1 \in E_i\}$ can be regarded as samples from a multinomial distribution, represented by

$$\mathbf{Z}_{ti} \mid \theta_i, m_{ti}^{\text{out}} \sim \text{Multinomial}(\mathbf{Z}_{ti} \mid \theta_i, m_{ti}^{\text{out}}), \quad (6.1)$$

where m_{ti}^{out} is a noise-free variable for y_{ti}^{out} , and where $m_{ti}^{\text{out}} = \sum_{j \in E_i} z_{tij}$. Since the transition populations $\mathbf{Z}$ are not observed, we treat them as latent variables.

We introduce noisy observation models of $\mathbf{Y}^{\text{out}}$ and $\mathbf{Y}^{\text{in}}$ with aggregation processes that represent flow conservation constraints. Regarding y_{ti}^{out} and y_{ti}^{in} as real numbers, we assume that y_{ti}^{out} and y_{ti}^{in} are generated from Gaussian distributions^{1 2},

$$y_{ti}^{\text{out}} \mid \mathbf{Z}, \sigma_i^2 \sim \mathcal{N}(y_{ti}^{\text{out}} \mid m_{ti}^{\text{out}}, \sigma_i^2), \quad (6.2)$$

$$y_{ti}^{\text{in}} \mid \mathbf{Z}, \eta_i^2 \sim \mathcal{N}(y_{ti}^{\text{in}} \mid m_{ti}^{\text{in}}, \eta_i^2), \quad (6.3)$$

where m_{ti}^{in} is a noise-free variable for y_{ti}^{in} . The transition populations $\mathbf{Z}$ satisfies the relations

$$m_{ti}^{\text{out}} = \sum_{j \in E_i} z_{tij}, \quad (6.4)$$

$$m_{ti}^{\text{in}} = \sum_{j \in E_i} \sum_{t'=1}^t W(\Delta_{tt'}; \pi_{ji}) z_{t'ji} \quad (6.5)$$

to ensure flow conservation laws. Equation (6.4) represents the aggregation constraints for outgoing counts, which states that the sum of people leaving location i at time step t equals the outgoing count at the same time step. Equation (6.5) represents the constraints for incoming counts, which states that the weighted sum of people leaving location j before time step t toward location i equals the incoming count of location i at time step t ; the idea behind (6.5) is that people who leave one location in one time step arrive at another location after some delay. Here, $\Delta_{tt'} = t - t'$ is travel duration, where t and t' are the arriving and leaving time steps, respectively. Weight function $W(\Delta; \pi_{ji})$ is a travel duration probability, where $\Delta \in \{0, \dots, T-1\}$, which is the probability that the travel duration is Δ ; π_{ji} denotes its parameters for transitions from location j to i , which are assumed time-independent, that is, the travel duration distributions do not vary over time. Details of $W(\Delta; \pi_{ji})$ are given in the following paragraphs. σ_i^2 in (6.2) and η_i^2 in (6.3) are noise variances for outgoing and incoming counts, respectively, at location i . The observation models (6.2) and (6.3) enable us to support noisy settings, which is the case where the flow conservation constraints (6.4) and 6.5 might not strictly hold. The noise can be regarded as factors that may disturb flow conservation laws, for example, people may enter or exit from the predefined space of interest (e.g., an exhibition hall).

We describe the formulation of travel duration probability $W(\Delta; \pi_{ji})$. We first introduce travel duration distribution $w(\tau; \pi_{ji})$ as the probability density function of continuous

¹The use of Gaussian noise models is advantageous in that we can use continuous optimization methods to estimate the transition populations. Details are described in Section 6.4.2.

²Although one could opt for Poisson noise models for y_{ti}^{out} and y_{ti}^{in} , it makes it difficult to estimate the degree of noise, for the variances of the Poisson distributions are determined by their means.

travel duration $\tau \geq 0$. The travel duration probability $W(\Delta; \pi_{ji})$ is defined by the following integral of $w(\tau; \pi_{ji})$ over the interval $[\Delta, \Delta + 1]$:

$$W(\Delta; \pi_{ji}) = \int_{\Delta}^{\Delta+1} w(\tau; \pi_{ji}) d\tau. \quad (6.6)$$

One can use any distribution function as $w(\tau; \pi_{ji})$, provided that $\int_0^{\infty} w(\tau; \pi_{ji}) d\tau = 1$.

The above formulation is inspired by the idea of describing diffusion processes of a piece of information on social networks with delays over time (Rodriguez, Balduzzi, and Schölkopf, 2011; Tanaka et al., 2016). Table 6.2 summarizes the travel duration distributions and the travel duration probabilities when using exponential, Rayleigh, and Weibull distributions, which are widely used for assessing duration in various fields such as user modeling (Wang et al., 2008) and diffusion modeling (Rodriguez, Balduzzi, and Schölkopf, 2011). The Weibull distribution is a more flexible distribution that has two parameters, where we set $\pi_{ji} = \{\alpha_{ji}, \beta_{ji}\}$. Here, $\alpha_{ji} > 0$ is the scale parameter and $\beta_{ji} > 0$ is the shape parameter of the distribution. The exponential distribution is a special case of the Weibull distribution with $\beta_{ji} = 1$; the Weibull distribution with $\beta_{ji} = 2$ is equivalent to the Rayleigh distribution. Figure 6.2 shows the examples of using the respective distributions.

In the formulation (6.6), travel durations are treated as random variables that follow the probability distributions with a few parameters; it is not a point estimate, which allows us to capture the heterogeneity in travel duration among individuals, efficiently.

We summarize the parameters of T-CFDM as follows: Transition probabilities $\theta = \{\theta_i \mid i \in V\}$; parameters of travel duration distributions $\pi = \{\pi_{ij} \mid i \in V; j \in E_i\}$; noise variances for outgoing counts $\sigma = \{\sigma_i \mid i \in V\}$; noise variances for incoming counts $\eta = \{\eta_i \mid i \in V\}$.

TABLE 6.2: Travel duration probabilities.

	Travel duration distribution $w(\tau; \pi_{ji})$		Travel duration probability $W(\Delta; \pi_{ji})$
Exponential	$\begin{cases} \alpha_{ji} e^{-\alpha_{ji}\tau} & \text{if } 0 \leq \tau \\ 0 & \text{otherwise} \end{cases}$		$e^{-\alpha_{ji}\Delta} - e^{-\alpha_{ji}(\Delta+1)}$
Rayleigh	$\begin{cases} \alpha_{ji} \tau e^{-\frac{1}{2}\alpha_{ji}\tau^2} & \text{if } 0 \leq \tau \\ 0 & \text{otherwise} \end{cases}$		$e^{-\frac{1}{2}\alpha_{ji}(\Delta)^2} - e^{-\frac{1}{2}\alpha_{ji}(\Delta+1)^2}$
Weibull	$\begin{cases} \alpha_{ji} \beta_{ji} (\alpha_{ji} \tau)^{\beta_{ji}-1} e^{-(\alpha_{ji} \tau)^{\beta_{ji}}} & \text{if } 0 \leq \tau \\ 0 & \text{otherwise} \end{cases}$		$e^{-(\alpha_{ji} \Delta)^{\beta_{ji}}} - e^{-(\alpha_{ji} (\Delta+1))^{\beta_{ji}}}$

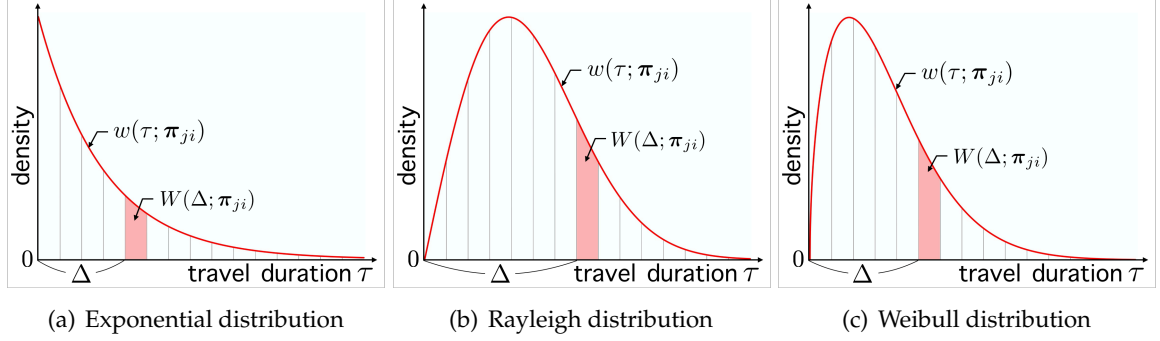


FIGURE 6.2: Visualization of travel duration probabilities. The red lines are the travel duration distributions $w(\tau; \pi_{ji})$; the red areas are the travel duration probabilities $W(\Delta; \pi_{ji})$. The gray vertical lines are spaced by unit time steps.

6.4.2 Inference

Outline

We describe an approximate expectation-maximization (EM) algorithm that can estimate the transition populations $\mathbf{Z}$ and the set of parameters $\Phi = \{\theta, \pi, \sigma, \eta\}$, simultaneously. Consider the estimation of Φ based on the EM algorithm (Bishop, 2006). Define $Q(\hat{\Phi}, \Phi)$ as the Q -function,

$$Q(\hat{\Phi}, \Phi) = \sum_{\mathbf{Z}} p(\mathbf{Z} | \mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, \hat{\Phi}) \log p(\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, \mathbf{Z} | \Phi), \quad (6.7)$$

which is the expectation of the logarithm of the perfect-data likelihood with respect to the posterior distribution of $\mathbf{z}$ conditioned on data $\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}$, and evaluated using the current parameter estimate $\hat{\Phi}$. Summing up over all possible values of $\mathbf{Z}$ in (6.7) is infeasible due to computation complexity, and thus we replace the posterior expectation with a *plug-in* procedure of the maximum a posteriori (MAP) estimate. The approximate Q -function is then obtained by plugging the MAP estimate $\hat{\mathbf{Z}}$ of $\mathbf{Z}$ represented by

$$\hat{\mathbf{Z}} = \arg \max_{\mathbf{Z}} \log p(\mathbf{Z} | \mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, \hat{\Phi}) \quad (6.8)$$

into the perfect-data log likelihood, as

$$Q^{\text{approx}}(\hat{\Phi}, \Phi) = \log p(\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, \hat{\mathbf{Z}} | \Phi). \quad (6.9)$$

A similar approximate EM algorithm was adopted in (Sheldon et al., 2013), and their experiments have shown that it provides an efficient and effective way of estimating the model parameters. In E-step, the MAP estimates of $\mathbf{Z}$ are obtained in (6.8). In M-step,

Algorithm 2: Inference procedure for T-CFDM.

Input : $\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, V, \{E_i \mid i \in V\}, T$ **Output:** $\mathbf{Z}, \theta, \pi, \sigma, \eta$

- 1: Initialize $\mathbf{Z}, \theta, \pi, \sigma, \eta$
 - 2: **repeat**
 - 3: /* E-step */
 - 4: Update $\mathbf{Z}$ by solving (6.11)
 - 5: /* M-step */
 - 6: Update θ_{ij} by (6.14) for $i \in V$; for $j \in E_i$
 - 7: Update π, σ, η by solving (6.15)
 - 8: **until** Convergence
-

the maximum likelihood estimates of Φ by maximizing $Q^{\text{approx}}(\hat{\Phi}, \Phi)$ (6.9). Algorithm 2 shows the inference procedure; details of E- and M-steps are provided in the following paragraphs.

E-step

Given the observations $\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}$ and the current estimate $\hat{\Phi}$, the log-posterior distribution of $\mathbf{Z}$ is given by

$$\begin{aligned}
 \log p(\mathbf{Z} \mid \mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, \hat{\Phi}) &\propto \log p(\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, \mathbf{Z} \mid \hat{\Phi}) \\
 &\approx \sum_{t=1}^T \sum_{i \in V} \left[-\frac{1}{2\hat{\sigma}_i^2} \left(y_{ti}^{\text{out}} - \sum_{j \in E_i} z_{tij} \right)^2 - \frac{1}{2\hat{\eta}_i^2} \left(y_{ti}^{\text{in}} - \sum_{j \in E_i} \sum_{t'=1}^t W(\Delta_{tt'}; \hat{\pi}_{ji}) z_{t'ji} \right)^2 \right. \\
 &\quad \left. + \left(\sum_{j \in E_i} z_{tij} \right) \log \left(\sum_{j \in E_i} z_{tij} \right) + \sum_{j \in E_i} \left(z_{tij} \log \hat{\theta}_{ij} - z_{tij} \log z_{tij} \right) \right] \\
 &=: \mathcal{L}_1(\mathbf{Z}),
 \end{aligned} \tag{6.10}$$

where we use Bayes' rule and the relations (6.4) and (6.5); we employ Stirling's approximation, $\log n! \approx n \log n - n$, in order to calculate $\log z_{tij}!$ and $(\sum_{j \in E_i} z_{tij})!$ efficiently, as in (Sheldon et al., 2013). Here, we define the objective function for $\mathbf{Z}$ as $\mathcal{L}_1(\mathbf{Z})$. The first and second terms in (6.10) show that the flow conservation laws are treated as the soft constraints; the noise variances σ_i^2, η_i^2 play a role in controlling the penalty incurred in violating the soft constraints. Finally, we arrive at the following optimization problem,

$$\begin{aligned}
 &\underset{\mathbf{Z}}{\text{minimize}} && -\mathcal{L}_1(\mathbf{Z}) \\
 &\text{subject to} && z_{tij} \geq 0, \quad t = 1, \dots, T; \quad i \in V; \quad j \in E_i,
 \end{aligned} \tag{6.11}$$

where we use a continuous relaxation from $z_{tij} \in \{0, 1, \dots, y_{ti}^{\text{out}}\}$ to $z_{tij} \geq 0$, which enables us to use various continuous optimization methods such as the L-BFGS-B method (Byrd et al., 1995). The first derivative of the objective function $\mathcal{L}_1(\mathbf{Z})$ (6.10) with respect to z_{tij} is given by

$$\begin{aligned} \frac{\partial \mathcal{L}_1(\mathbf{Z})}{\partial z_{tij}} = & \frac{1}{\hat{\sigma}_i^2} \left(y_{ti}^{\text{out}} - \sum_{j \in E_i} z_{tij} \right) + \frac{1}{\hat{\eta}_j^2} \sum_{k=1}^T \left(y_{kj}^{\text{in}} - \sum_{i \in E_j} \sum_{t=1}^k W(\Delta_{kt}; \hat{\pi}_{ij}) z_{tij} \right) W(\Delta_{kt}; \hat{\pi}_{ij}) \\ & + \log \left(\sum_{j \in E_i} z_{tij} \right) + \log \theta_{ij} - \log z_{tij}, \end{aligned} \quad (6.12)$$

which is required for estimating the transition populations $\mathbf{Z}$ using the L-BFGS-B method. Here, k is an auxiliary time variable.

M-step

Given the observations $\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}$ and the current MAP estimate $\hat{\mathbf{Z}}$, the approximate Q-function (6.9) is given by

$$\begin{aligned} Q^{\text{approx}}(\hat{\Phi}, \Phi) = & \log p(\mathbf{Y}^{\text{out}}, \mathbf{Y}^{\text{in}}, \hat{\mathbf{Z}} \mid \Phi) \\ \propto & \sum_{t=1}^T \sum_{i \in V} \left[-\frac{1}{2} \log \sigma_i^2 - \frac{1}{2} \sigma_i^2 \left(y_{ti}^{\text{out}} - \sum_{j \in E_i} \hat{z}_{tij} \right)^2 - \frac{1}{2} \log \eta_i^2 \right. \\ & \left. - \frac{1}{2\eta_i^2} \left(y_{ti}^{\text{in}} - \sum_{j \in E_i} \sum_{t'=1}^t W(\Delta_{tt'}; \pi_{ji}) \hat{z}_{t'ji} \right)^2 + \sum_{j \in E_i} \hat{z}_{tij} \log \theta_{ij} \right] \\ =: & \mathcal{L}_2(\theta, \pi, \sigma, \eta), \end{aligned} \quad (6.13)$$

where we define the objective function for the model parameters as $\mathcal{L}_2(\theta, \pi, \sigma, \eta)$. The maximum likelihood estimate of θ_{ij} is given by the closed-form solution via the Lagrangian multiplier method, represented by

$$\theta_{ij} = \frac{\sum_{t=1}^T \hat{z}_{tij}}{\sum_{t=1}^T \sum_{j \in E_i} \hat{z}_{tij}}. \quad (6.14)$$

The other parameters are estimated by maximizing the objective function (6.13). If one uses the Weibull distribution as the travel duration distribution that has two parameters, i.e., $\pi_{ji} = \{\alpha_{ji}, \beta_{ji}\}$, for each pair of locations, then the optimization problem to be solved

for the parameters π , σ , and η is as follows:

$$\begin{aligned} & \underset{\pi, \sigma, \eta}{\text{minimize}} && -\mathcal{L}_2(\theta, \pi, \sigma, \eta) \\ & \text{subject to} && \alpha_{ji} > 0, \quad j \in V; i \in E_j, \\ & && \beta_{ji} > 0, \quad j \in V; i \in E_j, \end{aligned} \tag{6.15}$$

which we solve by using the L-BFGS-B method (Byrd et al., 1995).

6.5 Experiments

6.5.1 Data

We used multiple real-world data sets: pedestrian data from exhibition halls, and bike trip data and taxi trip data from New York City, to demonstrate the effectiveness of the proposed model. We describe the details of the data sets in the following paragraphs.

Pedestrian data from exhibition halls

The data consist of pedestrian location logs in exhibition halls, which were collected at an event that attracted large crowds, Niconico Chokaigi 2016³, held at Makuhari Messe located near Tokyo, Japan, from 10:00 AM to 6:00 PM on April 29th, 2016. The event was spread over four exhibition halls, Hall 1, Hall 2, Hall 3, and Hall 4, the sizes of which are 186.3 m $\times$ 124.8 m, 183.2 m $\times$ 124.8 m, 127.6 m $\times$ 124.8 m, and 190.9 m $\times$ 108.3 m, respectively. The numbers $|V|$ of event booths in Hall 1, Hall 2, Hall 3, and Hall 4 were 38, 27, 10, and 9, respectively. It should be noted that we analyze the data for these four halls separately; namely we do not consider edges across halls but those within each hall only. E_i is a set of edges assuming a complete graph for each of the exhibition halls.

The pedestrian location logs were gathered by Bluetooth beacons installed at each booth, whose observation range is limited (at most 10 – 15 meters); it allows us to observe the times at which each pedestrian entered or left the observation range. The data consist of 3,727 mobile users who agreed to provide detailed information of location over time. The original data contained time stamps of arrival and departure at booths for each user, allowing users to be tracked over time.

In the experiments, we created the incoming and outgoing pedestrian count data at each booth, where the time interval was set to 3 minutes. Since the tendency of people flows,

³<http://www.chokaigi.jp/2016/en/>

TABLE 6.3: The totals of incoming and outgoing people in the real-world data sets.

Data	Pedestrian data							
Time-of-day	10:00 AM – 2:00 PM				2:00 PM – 6:00 PM			
Area	Hall 1	Hall 2	Hall 3	Hall 4	Hall 1	Hall 2	Hall 3	Hall 4
# outgoing	19,667	20,957	8,605	6,531	11,139	14,069	6,965	4,533
# incoming	19,840	21,198	8,770	6,630	10,974	13,829	6,806	4,437

Data	Bike trip data			
Time-of-day	8:00 AM – 4:00 PM		4:00 PM – 12:00 PM	
Date	Mar. 1	Jun. 1	Mar. 1	Jun. 1
# outgoing	7,275	13,277	7,091	14,931
# incoming	7,361	13,398	7,279	15,372

Data	Taxi trip data			
Time-of-day	8:00 AM – 4:00 PM		4:00 PM – 12:00 PM	
Date	Mar. 1	Jun. 1	Mar. 1	Jun. 1
# outgoing	93,514	91,911	107,923	104,188
# incoming	93,839	91,523	100,632	107,701

i.e., transition probabilities, could depend on time-of-day, the data were divided into two subsets, one from 10:00 AM to 2:00 PM, and the other from 2:00 PM to 6:00 PM; the number T of observation time steps was thus 80. The totals of incoming and outgoing pedestrians at all booths in each hall/time-of-day are shown in Table 6.3.

Traffic data in urban cities

We used bike trip data ⁴ and taxi trip data ⁵ in New York City, which are publicly available. These data sets consist of trip records of bikes/taxis, each record holds trip id, pickup location, dropoff location, pickup timestamp, and dropoff timestamp, where location information is available only when bikes/taxis started and finished their trips; the trajectories during the trips are not recorded.

In the experiments, we used the data from 8:00 AM to 4:00 PM and from 4:00 PM to 12:00 PM on March 1st and June 1st, and created the incoming and outgoing count data by

⁴<https://www.citibikenyc.com/system-data>

⁵http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

aggregating the original data into grid cells, the sizes of which in the bike trip data and the taxi trip data were $2 \text{ km} \times 2 \text{ km}$ (12×12 grid cells) and $3 \text{ km} \times 3 \text{ km}$ (18×18 grid cells), respectively. Here, we omitted grid cells whose incoming and outgoing counts were lower than a threshold. The number $|V|$ of grid cells for the bike trip data and the taxi trip data were then 11 and 14, respectively. The time interval in both data sets was set to 10 minutes; the number of observation time steps T was thus 48. E_i was a set of edges assuming a complete graph for both data sets. The totals of incoming and outgoing bikes/taxis at overall grid cells on each date/time-of-day are shown in Table 6.3.

6.5.2 Evaluation metric

Given the numbers of incoming and outgoing people $Y^{\text{in}}, Y^{\text{out}}$, we attempt to estimate transition populations Z between locations at each time step. The evaluation metric is the mean normalized absolute error (MNAE) in transition populations, represented by

$$\frac{1}{T} \sum_{t=1}^T \frac{\sum_{i \in V} \sum_{j \in E_i} |\hat{z}_{tij} - z_{tij}^*|}{\sum_{i \in V} \sum_{j \in E_i} z_{tij}^*}, \quad (6.16)$$

where z_{tij}^* is the true transition population and $\hat{z}_{tij}$ is its estimate.

6.5.3 Setup of T-CFDM

Different from the previous model (i.e., CFDM), T-CFDM can estimate transition populations and travel duration distributions, simultaneously. In our experiments, we used the Weibull distribution as the travel duration distribution between locations (See Section 6.4.1). The model parameters were learned by the approximate EM algorithm (described in Algorithm 2); the optimization problems were solved via the L-BFGS-B method (Byrd et al., 1995).

6.5.4 Baselines

We compared the proposed model, T-CFDM, with three baselines: *Uniform*, *Popularity*, and collective flow diffusion model (CFDM) (Kumar, Sheldon, and Srivastava, 2013), details of which are as follows.

- **Uniform.** We used a discrete uniform distribution to estimate transition populations. This baseline assumes that people move to neighbor locations with equal probability $1/|E_i|$, where $|E_i|$ is the number of neighbors of location i .

- **Popularity.** People are assumed to move to other locations in proportion to location popularity regardless of current locations; then the estimated transition population at time step t is given by

$$\hat{z}_{tij} = y_{ti}^{\text{out}} \times \frac{\sum_{t=1}^T y_{tj}^{\text{in}}}{\sum_{t=1}^T \sum_{j \in E_i} y_{tj}^{\text{in}}}. \quad (6.17)$$

- **CFDM.** This is the basic model of the proposed model. In CFDM, people are assumed to be complete their transitions within each time step; the transition populations $\mathbf{Z}$ satisfy the flow conservation laws,

$$m_{ti}^{\text{out}} = \sum_{j \in E_i} z_{tij}, \quad (6.18)$$

$$m_{ti}^{\text{in}} = \sum_{j \in E_i} z_{tji}. \quad (6.19)$$

The model parameters were learned by the approximate EM algorithm, as in the proposed model, where means of the observation models (6.2) and (6.3) were replaced with the relations (6.18) and (6.19).

6.5.5 Results

Transition population estimation

Table 6.4 shows MNAE and the standard error attained by T-CFDM, CFDM, Popularity, and Uniform. For all data sets, the MNAEs of T-CFDM were lower than the compared methods, and the differences between T-CFDM and the other methods were statistically significant (Student's t-test) at the level of $P < 0.01$. These results indicate that the introduction of the travel duration distributions is effective for accurately estimating the transition populations between locations. In the following, we present the results of further assessment of T-CFDM.

Travel duration probability estimation

We evaluated performance of estimating travel duration probabilities; the evaluation metric is mean Kullback-Leibler (KL) divergence between the true and the estimated travel

TABLE 6.4: Mean normalized absolute errors (MNAEs) and standard errors of the estimates of transition populations.

(a) Pedestrian data (10:00 AM – 2:00 PM)				
	Hall 1	Hall 2	Hall 3	Hall 4
T-CFDM	1.174 ± 0.016	0.995 ± 0.013	0.689 ± 0.014	0.600 ± 0.022
CFDM	1.261 ± 0.009	1.127 ± 0.010	0.864 ± 0.013	0.672 ± 0.022
Popularity	1.674 ± 0.008	1.500 ± 0.008	1.106 ± 0.016	0.779 ± 0.020
Uniform	1.767 ± 0.007	1.596 ± 0.007	1.130 ± 0.015	1.165 ± 0.016
(b) Pedestrian data (2:00 PM – 6:00 PM)				
	Hall 1	Hall 2	Hall 3	Hall 4
T-CFDM	1.439 ± 0.026	1.436 ± 0.010	0.908 ± 0.022	0.743 ± 0.043
CFDM	1.506 ± 0.025	1.515 ± 0.010	0.995 ± 0.026	0.780 ± 0.045
Popularity	1.825 ± 0.017	1.789 ± 0.010	1.253 ± 0.029	0.924 ± 0.038
Uniform	1.904 ± 0.016	1.857 ± 0.010	1.289 ± 0.029	1.282 ± 0.038
(c) Bike trip data				
	8:00 AM – 4:00 PM		4:00 PM – 12:00 PM	
	Mar. 1	Jun. 1	Mar. 1	Jun. 1
T-CFDM	0.561 ± 0.013	0.630 ± 0.010	0.628 ± 0.031	0.590 ± 0.027
CFDM	0.655 ± 0.012	0.737 ± 0.012	0.706 ± 0.029	0.651 ± 0.027
Popularity	0.691 ± 0.013	0.751 ± 0.012	0.739 ± 0.028	0.671 ± 0.027
Uniform	1.025 ± 0.009	0.953 ± 0.013	1.036 ± 0.025	0.943 ± 0.025
(d) Taxi trip data				
	8:00 AM – 4:00 PM		4:00 PM – 12:00 PM	
	Mar. 1	Jun. 1	Mar. 1	Jun. 1
T-CFDM	0.363 ± 0.004	0.433 ± 0.004	0.429 ± 0.022	0.366 ± 0.012
CFDM	0.472 ± 0.004	0.496 ± 0.004	0.469 ± 0.059	0.448 ± 0.012
Popularity	0.505 ± 0.004	0.520 ± 0.004	0.477 ± 0.016	0.493 ± 0.012
Uniform	1.043 ± 0.005	1.043 ± 0.005	0.923 ± 0.014	0.929 ± 0.010

TABLE 6.5: Mean Kullback-Leibler divergence for the estimation of travel duration probability.

(a) Pedestrian data (10:00 AM – 2:00 PM)				
	Hall 1	Hall 2	Hall 3	Hall 4
T-CFDM	1.567 ± 0.050	1.547 ± 0.050	1.627 ± 0.087	1.669 ± 0.231
CFDM	1.800 ± 0.091	1.695 ± 0.082	1.675 ± 0.122	1.748 ± 0.127
(b) Pedestrian data (2:00 PM – 6:00 PM)				
	Hall 1	Hall 2	Hall 3	Hall 4
T-CFDM	1.583 ± 0.051	1.560 ± 0.053	1.613 ± 0.071	1.560 ± 0.125
CFDM	1.758 ± 0.078	1.737 ± 0.103	1.670 ± 0.087	1.854 ± 0.082
(c) Bike trip data				
	8:00 AM – 4:00 PM		4:00 PM – 12:00 PM	
	Mar. 1	Jun. 1	Mar. 1	Jun. 1
T-CFDM	1.489 ± 0.194	1.353 ± 0.172	1.498 ± 0.218	1.182 ± 3.129
CFDM	2.936 ± 0.048	3.052 ± 0.047	2.942 ± 0.056	3.129 ± 0.043
(d) Taxi trip data				
	8:00 AM – 4:00 PM		4:00 PM – 12:00 PM	
	Mar. 1	Jun. 1	Mar. 1	Jun. 1
T-CFDM	1.432 ± 0.142	2.444 ± 0.135	2.060 ± 0.161	2.130 ± 0.102
CFDM	3.123 ± 0.058	3.069 ± 0.049	3.044 ± 0.072	3.062 ± 0.058

duration probabilities over respective pairs of locations, represented by

$$\frac{1}{|V|} \sum_{j \in V} \frac{1}{|E_j|} \sum_{i \in E_j} \sum_{\Delta=0}^{T-1} P_{ji}^*(\Delta) \log \frac{P_{ji}^*(\Delta)}{W(\Delta; \hat{\pi}_{ji})}, \quad (6.20)$$

where $P_{ji}^*(\Delta)$ and $W(\Delta; \hat{\pi}_{ji})$ are the true travel duration probability and its estimate, respectively, for transition from location j to location i . It should be noted that we can use the true probability $P_{ji}^*(\Delta)$ for evaluations because the original data contain the tracking information with user identifiers (See 6.5.1). Table 6.5 shows the mean KL divergence and the standard error for T-CFDM and CFDM, where CFDM assumed that all people who leave a location at a time step have always arrived at another location in the same time step; the travel duration probability $W(\Delta; \hat{\pi}_{ji})$ equals 1 if $\Delta = 0$ and 0 otherwise. As shown in Table 6.5, T-CFDM yielded lower mean KL divergence values for all data sets. The results show that T-CFDM can accurately estimate the travel duration probabilities.

The improvements of bike trip data and taxi trip data were larger than those of the pedestrian data, which is reasonable because the distances between locations (i.e., grid cells) were relatively large in the urban bikes/taxis trip data compared with the pedestrian data in the exhibition halls; in such cases, the travel duration probability is more helpful. The results of Tables 6.4 and 6.5 indicate that the incorporation of the people's travel durations into the model is important for estimating the transition populations accurately.

Figure 6.3 illustrates visualization examples of transition matrices and travel duration probabilities estimated by T-CFDM for the respective times-of-day in the bike trip data and the taxi trip data. The transition matrix is the total number of bikes/taxis that moved between each pair of origin location i and destination location j , whose elements in the transition matrix were calculated as follows: $\hat{z}_{ij} = \sum_{t=1}^T \hat{z}_{tij}$. As shown in Figure 6.3, T-CFDM accurately estimated transition matrix from the numbers of incoming and outgoing people. One also observes that T-CFDM flexibly fitted to the various travel duration distributions. The results show that T-CFDM allows us to capture the heterogeneity in travel duration among individuals as it can handle the travel durations as the random variables; it does not use the point estimates of travel durations.

Visualization of the estimated people flows

We finally present qualitative comparisons of the estimated transition populations in the pedestrian data. Figure 6.4 visualizes the pedestrian flows between locations (i.e., booths) in each hall, where we illustrate the true pedestrian flow on the left, and the estimates of T-CFDM, CFDM, Popularity and Uniform on the right. As shown in Figure 6.4, T-CFDM better discerned the pedestrian flows than the other methods. CFDM tended to output some false flows. This is mainly because CFDM is based on the unrealistic assumption that all pedestrians who left a location in one time step should arrive at another location in the same time step; and thus CFDM mis-estimated the transition populations between booths. T-CFDM, on the other hand, could more accurately estimate the transition populations because it considers travel durations between booths.

The visualization results are useful for optimizing navigation systems. For example, discovering popular routes of pedestrians yields better route recommendations; those that are likely to be chosen by visitors. The results are also useful to marketers when they want to optimize the strategies of location-based advertising. For example, analyzing the transition relation between booths provides better understanding of the visitors' interests. This makes it easier for marketers to determine which advertisements to serve to the visitors according to their interests and current locations.

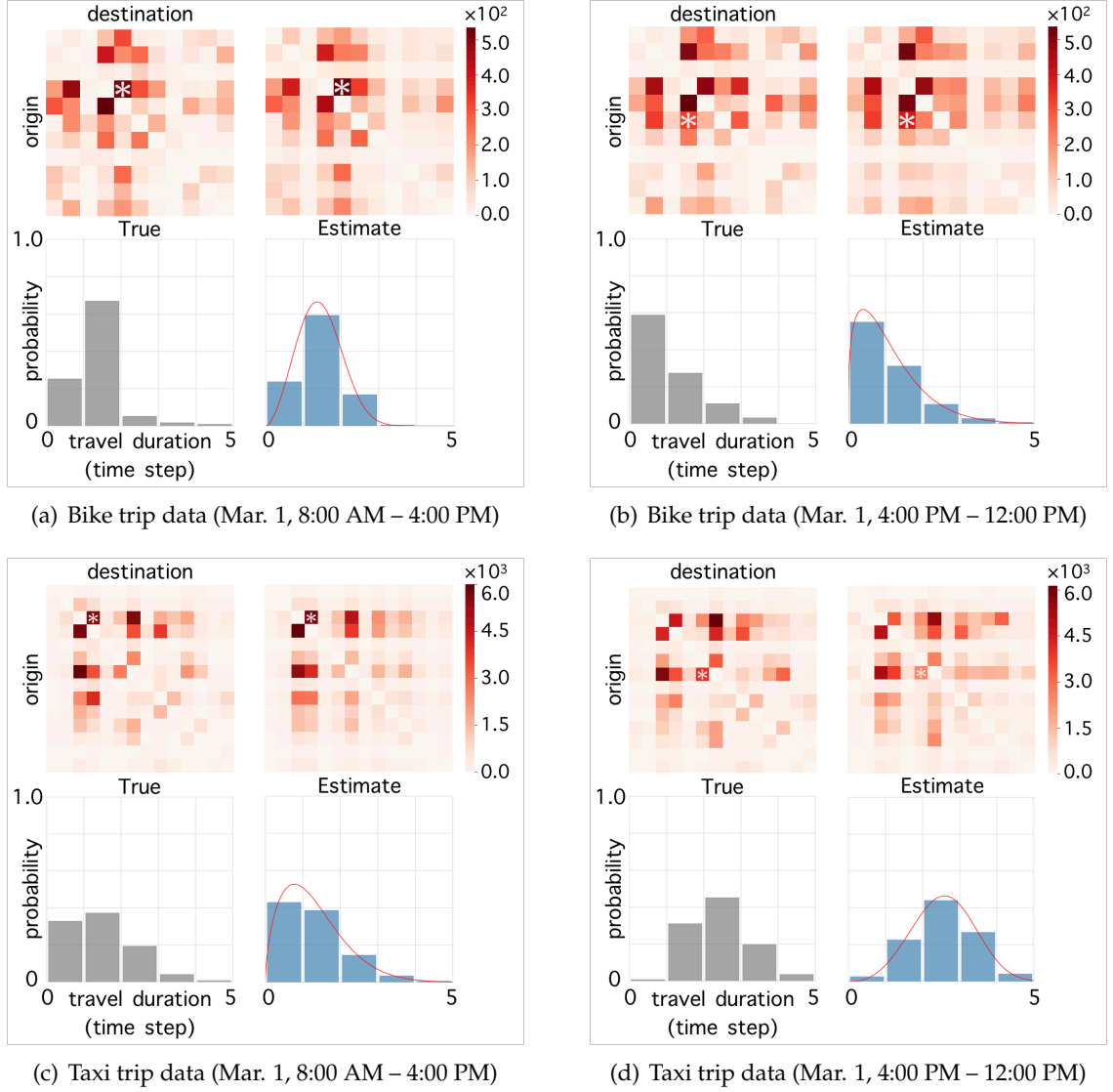
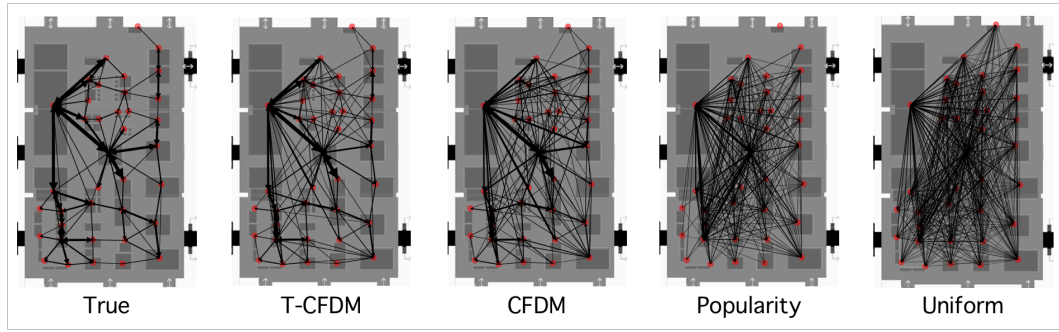
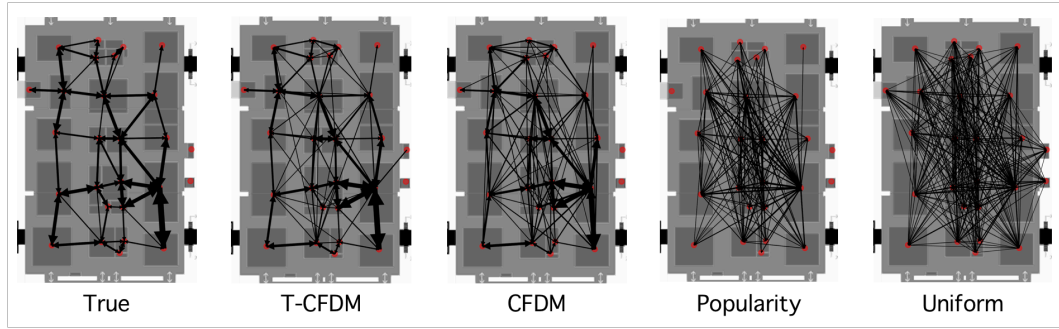


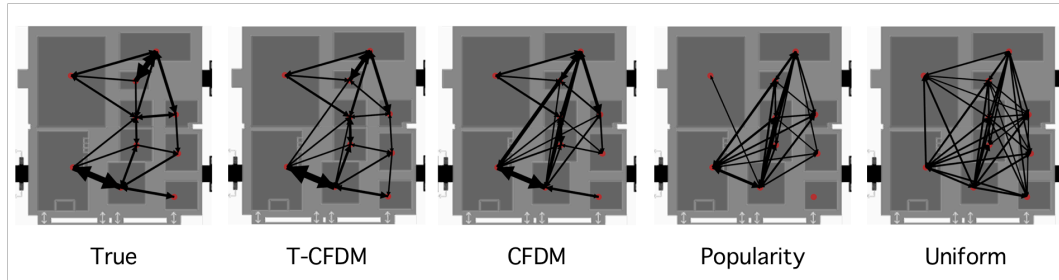
FIGURE 6.3: Visualization of transition matrices and travel duration probabilities. (Upper part of each figure) Heatmap visualization of the transition matrix. The true matrix is shown on the left and the estimate of T-CFDM is shown on the right. (Lower part of each figure) Travel duration probability $W(\Delta; \hat{\pi}_{ji})$ estimated by T-CFDM for the origin-destination pair specified by the white asterisk (*) in the corresponding upper heatmap; the left one is the true probability and the right one is the estimate of T-CFDM. Red line represents the travel duration distribution $w(\tau; \hat{\pi}_{ji})$ estimated by T-CFDM.



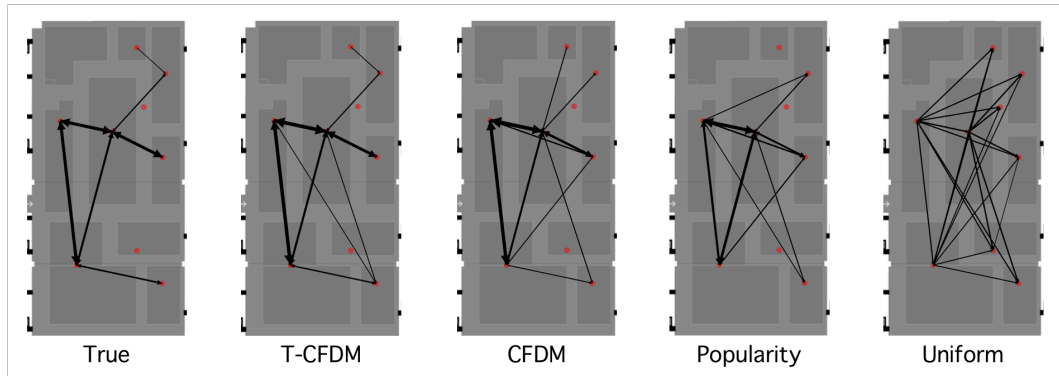
(a) Hall 1 (10:00 AM – 2:00 PM)



(b) Hall 2 (10:00 AM – 2:00 PM)



(c) Hall 3 (10:00 AM – 2:00 PM)



(d) Hall 4 (10:00 AM – 2:00 PM)

FIGURE 6.4: Comparison of pedestrian flows between booths in each hall. The red dots represent the locations of booths, and the directed edges represent the pedestrian flows between booths. The edge widths are proportional to the transition populations between the respective booth pairs. Note that we omitted those edges whose transition populations were lower than a threshold, and bidirectional edge widths are proportional to the average of the transition populations between the pair of locations.

6.6 Discussion

In this chapter, we proposed time-delayed collective flow diffusion models (T-CFDM) that can estimate latent people flows, i.e., transition populations between locations, from just aggregated population data. We designed an observation model with aggregation processes, which represent flow conservation constraints. A key idea is to incorporate travel duration probabilities into the aggregation constraints. We also developed the approximate expectation-maximization algorithm, which enables us to learn the transition populations and the model parameters simultaneously.

The experiments on multiple real-world data sets showed that T-CFDM can 1) accurately infer transition populations between locations and 2) flexibly and precisely capture the travel duration probabilities between locations.

There are several important research issues that should be considered in the future. First, we could extend the model to capture time-varying people flows. The previous study has shown that the models taking account of flow changes over time achieve the improvements of estimation performance (Iwata et al., 2017). Second, it would be interesting to investigate advanced inference algorithms. In this thesis, we used a continuous relaxation for transition populations $\mathbf{Z}$ and solved the problem (6.11) using continuous optimization methods. Another option is to directly handle discrete variables $\mathbf{Z}$ in inference processes; the recent work has shown that one can formulate the maximum a posteriori problem for $\mathbf{Z}$ in the original collective flow diffusion model as a combinatorial optimization problem, called (non-linear) minimum cost flow problem (Akagi et al., 2020 (to appear)). The idea in this work would be beneficial for constructing an efficient and effective alternative for learning transition populations in T-CFDM inference procedures. Finally, it would be a fruitful direction to utilize additional information such as weather conditions and satellite images of the city by combining it with non-linear models (e.g., deep neural networks), as in (Iwata and Shimizu, 2019); it could be expected to provide performance improvements.

Part III

Conclusion and Future Directions

Chapter 7

Conclusion

7.1 Summary of contributions

Location data are now being accumulated in various spaces such as cities and exhibition halls. The utilization of them is important for improving city environments and location-based services. Location data are, however, often aggregated over spatial regions due to practical reasons (e.g., privacy concerns), which makes it difficult to apply existing statistical methods for spatial data analysis. In this thesis, we explored a framework that can yield prediction and knowledge discovery from spatially aggregated data.

Specifically, we considered two fundamental and important problems: 1) refining coarse-grained aggregated data (Part I); 2) inferring latent people flow from aggregated population data (Part II). Our solution is the probabilistic models that incorporate spatial aggregation processes, a critical functionality for learning the models from aggregated data. We also developed the efficient inference procedure for each proposed model. We confirmed the effectiveness of the proposed models on the basis of experiments using real-world data sets. The contributions of each part are summarized as follows.

Refining coarse-grained aggregated data

The aim of Part I was to construct a framework based on Gaussian processes (GPs) for predicting fine-grained data via the use of multiple aggregated data sets with various granularities. The two questions to be answered were

- *How do we model the generative process of data aggregated over spatial regions?*
- *How do we establish dependences between multiple aggregated data sets?*

In order to answer the first question, Chapter 3 formulated a probabilistic model, called Spatially Aggregated Gaussian Processes with a Single Output (SAGP-S), that can infer a

function defined on the continuous space from aggregated observations. A key technique is introducing an observation model with aggregation processes, which is represented by the integral of the GP. This formulation is advantageous in that variances and covariances can be evaluated considering sizes and shapes of regions. SAGP-S is a core idea for constructing the proposed models in Chapters 4 and 5.

For the second question, we examined two types of methodologies that can introduce dependences between data sets: 1) regression approach and 2) multivariate approach. In Chapter 4, we proposed a new regression-based model, called Two-stage Spatially Aggregated Gaussian Process Model (2-stage SAGP model), in which GP for target data is assumed to be a linear combination of the posterior GPs for auxiliary data sets. One benefit of 2-stage SAGP model is that the regression coefficients are determined by not only the strength of relationships with the target data but also the posterior variances for each auxiliary data set. In Chapter 5, we proposed a novel multivariate GP model, called Spatially Aggregated Gaussian Processes with Multiple Outputs (SAGP-M), that can infer multivariate function from multiple aggregated data sets with various granularities; this is a natural extension of SAGP-S. In SAGP-M, multivariate dependent GP for multiple data sets is defined by a linear mixing of independent latent GPs, which can be expected to accurately estimate spatial correlations even if some of input data sets have a coarse granularity. The sharing of latent GPs is also advantageous in that it can support data sets from multiple domains (e.g., cities).

We also developed the Bayesian inference procedures for 2-stage SAGP model and SAGP-M. The effectiveness of each model was demonstrated through experiments on real-world aggregated data sets from urban cities. As a result, we confirmed that SAGP-M yielded better refinement performance than the compared methods; 2-stage SAGP model, on the other hand, has the advantage in terms of the computation costs compared with multivariate approaches.

Inferring latent people flow from aggregated population data

The task of Part II was to capture people's transitions from just aggregated population data at each observed location. We proposed a probabilistic model, called Time-delayed Collective Flow Diffusion Model (T-CFDM), that can estimate transition populations between locations given the numbers of incoming and outgoing people at each location and at each time step. In T-CFDM, people's transition is modeled as diffusion processes that are governed by time-independent and location-dependent transition probabilities. By introducing an observation model with aggregation processes that represent flow conservation constraints, one can estimate transition populations from just aggregated population data. Different from the existing models, we incorporated travel duration probabilities

into the aggregation constraints, which allows the accurate estimation of transition populations in more practical situations, that is, the observation ranges are limited and some people are not observed in any location in some periods. We also developed the approximate expectation-maximization algorithm that can learn the transition populations and the model parameters, simultaneously. Through the experiments on real-world data sets, we demonstrated that T-CFDM outperformed the existing models in terms of transition population estimation performance.

7.2 Future directions

This thesis explored probabilistic approaches for modeling spatially aggregated data, which is a practically and theoretically important research topic. The results presented in this thesis point to several interesting future directions, some of which we shall outline from a global perspective, thus concluding the thesis.

Incorporating rich additional information

One promising direction for advanced modeling of spatially aggregated data is to incorporate a wide variety of rich additional information: event programs, weather conditions, road networks, satellite images, and so on. Most of them take the form of unstructured data such as images and texts. A powerful option to leverage such data is to use deep neural networks as a model component, which can automatically extract the meaningful feature representations from unstructured data. It would be interesting to develop a framework for integrating neural networks into the model and learning their parameters in aggregate settings, efficiently.

Knowledge transfer across domains

Although aggregated data are now being accumulated in the world, its data collection is still limited. For example, one could be relatively easy to obtain various kinds of data sets from an urban city (e.g., New York City); however, one could not have few data sets from rural and/or developing areas. As another example, large-scale events (e.g., exhibitions) are rarely held, then it might be difficult to gather the large amount of data under the same conditions. In such situations, a transfer learning approach has promise, because it can be used for sharing knowledge derived from data sets in multiple domains including cities or exhibition halls. This might help train the models in the case where we have a few data sets in some domains. To my best knowledge, transfer learning methods in aggregate

settings have not been well-established. We believe that an exploration of a framework for integrating various kinds of data sets in various domains will open up new research directions.

Hybrid approach

Another important research direction is to construct a hybrid approach that can utilize both individual-level data and aggregate-level data. In practical situations, one could use the small amount of individual-level data, the collection of which is time-consuming and costly. It would be interesting to develop a framework for learning the model of scarce individual-level data by utilizing various kinds of aggregate-level data sets. If the attribute and/or domain of scarce individual-level data is unseen, namely, not included in the training data sets, solving this problem is challenging because it is not straightforward to know the relationships between scarce individual-level data and aggregate-level data sets. We believe that the solution to this problem would further increase the usefulness of aggregate-level data.

Appendix A

On integrability of the observation model in SAGP-S

In this appendix, we discuss conditions for the observation model (3.2) to be well defined. We assume that $X \subset \mathbb{R}^2$ is a bounded Jordan-measurable set, and that elements $a_n(\mathbf{x})$ of $\mathbf{a}(\mathbf{x})$ are Riemann integrable on X . The latter assumption is satisfied when $a_n(\mathbf{x})$ is defined as in (3.3) with Jordan-measurable regions R_n . It is known that a continuous function on X is Riemann integrable on X , and that a product of Riemann integrable functions is again Riemann integrable; thus, a sufficient condition for $\mathbf{a}(\mathbf{x})f(\mathbf{x})$ in the observation model (3.2) to be Riemann integrable is that the prior process $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ is sample-path continuous.

The assumption made in (3.2) of the integrability of $m(\mathbf{x})$ allows us to reduce integrability of the prior process $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ to sample-path continuity of the zero-mean process $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ on X . A sufficient condition (Adler and Taylor, 2007, Theorem 1.4.1) for the sample-path continuity of the zero-mean Gaussian process is that for some $0 < C < \infty$ and $\xi, \epsilon > 0$

$$k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2k(\mathbf{x}, \mathbf{x}') \leq \frac{C}{|\log \|\mathbf{x} - \mathbf{x}'\||^{1+\xi}} \quad (\text{A.1})$$

holds for all $\mathbf{x}, \mathbf{x}'$ with $\|\mathbf{x} - \mathbf{x}'\| < \epsilon$. If one uses the squared-exponential kernels for $k(\mathbf{x}, \mathbf{x}')$, then one can confirm that the above condition is satisfied, and consequently the observation model (3.2) is well defined.

It should be noted that the sample-path continuity discussed above is different from the mean-square (MS) continuity. A process $f(\mathbf{x})$ is said to be MS continuous at $\mathbf{x} = \mathbf{x}_*$ if for any sequence $\mathbf{x}_i$ converging to $\mathbf{x}_*$ as $i \rightarrow \infty$ it holds that $\mathbb{E}[\|f(\mathbf{x}_i) - f(\mathbf{x}_*)\|^2] \rightarrow 0$ as $i \rightarrow \infty$. A necessary and sufficient condition for a random field to be MS continuous at $\mathbf{x}_*$ is that its covariance function $k(\mathbf{x}, \mathbf{x}')$ is continuous at the point $\mathbf{x} = \mathbf{x}' = \mathbf{x}_*$ (Papoulis, 1991, Appendix 10A), which in the case of Gaussian processes is weaker than the above sufficient condition for the sample-path continuity.

Bibliography

Adler, R. J. and J. E. Taylor (2007). *Random Fields and Geometry*. Springer.

Akagi, Y., T. Nishimura, T. Kurashima, and H. Toda (2018). “A fast and accurate method for estimating people flow from spatiotemporal population data”. In: *IJCAI’18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3293–3300.

Akagi, Y., T. Nishimura, Y. Tanaka, T. Kurashima, and H. Toda (2020 (to appear)). “Exact and Efficient Inference for Collective Flow Diffusion Model via Minimum Convex Cost Flow Algorithm”. In: *AAAI’20: Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

Barlacchi, G., M. De Nadai, R. Larcher, A. Casella, C. Chitic, and j G. Torrisi et al. (2015). “A Multi-source Dataset of Urban Life in the City of Milan and the Province of Trentino”. In: *Scientific Data* 2.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland (2014). “Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data”. In: *ICMI’14: Proceedings of the 16th ACM International Conference on Multimodal Interaction*, pp. 427–434.

Bonilla, E., K. M. Chai, and C. Williams (2008). “Multi-task Gaussian process prediction”. In: *NeurIPS’08: Advances in Neural Information Processing Systems*, pp. 153–160.

Boyle, P. and M. Frean (2005). “Dependent Gaussian Processes”. In: *NeurIPS’05: Advances in Neural Information Processing Systems*, pp. 217–224.

Burgess, T. M. and R. Webster (1980). “Optimal Interpolation and Isarithmic Mapping of Soil Properties”. In: *Journal of Soil Science* 31.2.

Byrd, Richard H., P. Lu, J. Nocedal, and C. Zhu (1995). “A Limited Memory Algorithm for Bound Constrained Optimization”. In: *SIAM Journal on Scientific Computing* 16, pp. 1190–1208.

- Cannon, A. J. (2011). "Quantile Regression Neural Networks: Implementation in R and Application to Precipitation Downscaling". In: *Computers & Geosciences* 37.9, pp. 1277–1284.
- Chan, Antoni B., Z. -S. J. Liang, and N. Vasconcelos (2008). "Privacy Preserving Crowd Monitoring: Counting People without People Models or Tracking". In: *CVPR'08: IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7.
- City of Chicago. *Chicago Data Portal*. <https://data.cityofchicago.org/>.
- City of New York. *NYC Open Data*. <https://opendata.cityofnewyork.us/>.
- Dhar, S. and U. Varshney (2011). "Challenges and Business Models for Mobile Location-based Services and Advertising". In: *Communications of the ACM* 54.5, pp. 121–129.
- Diodato, N., G. Bellocchi, C. Bertolin, and D. Camuffo (2010). "Multiscale Regression Model to Infer Historical Temperatures in a Central Mediterranean Sub-regional Area". In: *Climate of the Past Discussions* 6, pp. 2625–2649. DOI: 10.5194/cpd-6-2625-2010.
- Dong, C., C. C. Loy, K. He, and X. Tang (2014). "Learning a Deep Convolutional Network for Image Super-resolution". In: *ECCV'14: Proceedings of the 13rd European Conference on Computer Vision*, pp. 184–199.
- Du, J., A. Kumar, and P. Varakantham (2014). "On Understanding Diffusion Dynamics of Patrons at a Theme Park". In: *AAMAS'14: Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1501–1502.
- Flaxman, S. R., Y. -X. Wang, and A. J. Smola (2015). "Who Supported Obama in 2012?: Ecological Inference Through Distribution Regression". In: *KDD'15: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 289–298.
- Ghosh, S. (2010). "SVM-PGSL Coupled Approach for Statistical Downscaling to Predict Rainfall from GCM Output". In: *Journal of Geophysical Research: Atmospheres* 115.D22.
- Giannotti, F., M. Nanni, F. Pinelli, and D. Pedreschi (2007). "Trajectory Pattern Mining". In: *KDD'07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 330–339.
- Gibbs, M. and D.J.C. MacKay (1997). "Efficient implementation of Gaussian processes". In: *Technical Report*. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.8224>.
- Goldstein, B. and L. Dyson (2013). *Beyond Transparency: Open Data and the Future of Civic Innovation*. Code for America Press.

- Goovaerts, P. (2010). "Combining areal and point data in geostatistical interpolation: Applications to soil science and medical geography". In: *Mathematical Geosciences* 42.5, pp. 535–554.
- Gotway, C. A. and L. J. Young (2002). "Combining Incompatible Spatial Data". In: *Journal of the American Statistical Association* 97.458, pp. 632–648.
- Hamelijnck, O., T. Damoulas, K. Wang, and M. Girolami (2019). "Multi-resolution Multi-task Gaussian Processes". In: *NeurIPS'19: Advances in Neural Information Processing Systems*, pp. 14025–14035.
- Hessami, M., P. Gachon, T. BMJ Ouarda, and André St-Hilair (2008). "Automated Regression-based Statistical Downscaling Tool". In: *Environmental Modeling & Software* 23.6, pp. 813–834.
- Higdon, D. (2002). "Space and space-time modelling using process convolutions". In: *Quantitative Methods for Current Environmental Issues*, pp. 37–56.
- Hoang, M., Y. Zheng, and A. Singh (2016). "FCCF: Forecasting Citywide Crowd Flows Based on Big Data". In: *SIGSPATIAL'16: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–10.
- Howitt, R. and A. Reynaud (2003). "Spatial disaggregation of agricultural production data using maximum entropy". In: *European Review of Agricultural Economics* 30.2, pp. 359–387.
- Huang, H. and G. Gartner (2010). "A Survey of Mobile Indoor Navigation Systems". In: *Cartography in Central and Eastern Europe*. Ed. by Georg Gartner and Felix Ortig. Lecture Notes in Geoinformation and Cartography. Berlin Heidelberg: Springer, pp. 305–319.
- Iwata, T., A. Shah, and Z. Ghahramani (2013). "Discovering Latent Influence in Online Social Activities via Shared Cascade Poisson Processes". In: *KDD'13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 266–274.
- Iwata, T. and H. Shimizu (2019). "Neural collective graphical models for estimating spatio-temporal population flow from aggregated data". In: *AAAI'19: Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 3935–3942.
- Iwata, T., H. Shimizu, F. Naya, and N. Ueda (2017). "Estimating People Flow from Spatio-temporal Population Data via Collective Graphical Mixture Models". In: *ACM Transactions on Spatial Algorithms and Systems* 3.1, pp. 1–18.

- Jerrett, M., R. T. Burnett, B. S. Beckerman, M. C. Turner, D. Krewski, G. Thurston, and et al. (2013). "Spatial Analysis of Air Pollution and Mortality in California". In: *American Journal of Respiratory and Critical Care Medicine* 188.5, pp. 593–599.
- Keil, P., J. Belmaker, A. M. Wilson, P. Unitt, and W. Jetz (2013). "Downscaling of species distribution models: a hierarchical approach". In: *Methods in Ecology and Evolution* 4.1, pp. 82–94.
- Kempe, D., J. Kleinberg, and É. Tardos (2003). "Maximizing the Spread of Influence through a Social Network". In: *KDD'03: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146.
- King, G. (2013). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press.
- Klein, Lawrence A. (2001). *Sensor Technologies and Data Requirements for ITS*. Artech House.
- Kumar, A., D. Sheldon, and B. Srivastava (2013). "Collective Diffusion over Networks: Models and Inference". In: *UAI'13: Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pp. 351–360.
- Kurashima, T., T. Iwata, G. Irie, and K. Fujimura (2010). "Travel Route Recommendation Using Geotags in Photo Sharing Sites". In: *CIKM'10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 579–588.
- Kurashima, T., T. Iwata, N. Takaya, and H. Sawada (2014). "Probabilistic Latent Network Visualization: Inferring and Embedding Diffusion Networks". In: *KDD'14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1236–1245.
- Law, H. C. L., D. Sejdinovic, E. Cameron, T. C. D. Lucas, S. Flaxman, K. Battle, and K. Fukumizu (2018). "Variational learning on aggregate outputs with Gaussian processes". In: *NeurIPS'18: Advances in Neural Information Processing Systems*, pp. 6084–6094.
- Leskovec, J., Lada A. Adamic, and Beuardo A. Huberman (2007). "The Dynamics of Viral Marketing". In: *ACM Transactions on the Web* 1.1.
- Liu, D. C. and J. Nocedal (1989). "On the Limited Memory BFGS Method for Large Scale Optimization". In: *Mathematical Programming* 45.1–3, pp. 503–528.
- Luttinen, J. and A. Ilin (2009). "Variational Gaussian-process factor analysis for modeling spatio-temporal data". In: *NeurIPS'09: Advances in Neural Information Processing Systems*, pp. 1177–1185.

- Micchelli, C. A. and M. Pontil (2004). "Kernels for multi-task learning". In: *NeurIPS'04: Advances in Neural Information Processing Systems*, pp. 921–928.
- Miller, B. A., S. Koszinski, M. Wehrhan, and M. Sommer (2015). "Impact of Multi-scale Predictor Selection for Modeling Soil Properties". In: *Geoderma* 239–240, pp. 97–106.
- Misra, S., S. Sarkar, and P. Mitra (2017). "Statistical Downscaling of Precipitation Using Long Short-term Memory Recurrent Neural Networks". In: *Theoretical and Applied Climatology*. URL: <https://doi.org/10.1007/s00704-017-2307-2>.
- Monreale, A., F. Pinelli, R. Trasarti, and F. Giannotti (2009). "WhereNext: A Location Predictor on Trajectory Pattern Mining". In: *KDD'09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 637–646.
- Murakami, D. and M. Tsutsumi (2011). "A New Areal Interpolation Technique Based on Spatial Econometrics". In: *Procedia-Social and Behavioral Sciences* 21, pp. 230–239.
- Murray-Smith, R. and B. A. Pearlmutter (2004). "Transformations of Gaussian Process Priors". In: *DSMML'04: Deterministic and Statistical Methods in Machine Learning, First International Workshop*, pp. 110–123.
- Myers, D. E. (1984). "Co-kriging – New Developments". In: *Geostatistics for Natural Resources Characterization: Part 1*. Ed. by G. Verly, M. David, A. G. Journel, and A. Marechal. Vol. 122. NATO ASI Series C: Mathematical and Physical Sciences. Dordrecht: D. Reidel Publishing, pp. 295–305.
- Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 3rd edition.
- Park, N. -W. (2013). "Spatial Downscaling of TRMM Precipitation Using Geostatistics and Fine Scale Environmental Variables". In: *Advances in Meteorology* 2013, pp. 1–9.
- Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Robinson, W. S. (1950). "Ecological correlations and the behavior of individuals". In: *American Sociological Review* 15, pp. 351–357.
- Rodriguez, Manuel G., D. Balduzzi, and B. Schölkopf (2011). "Uncovering the Temporal Dynamics of Diffusion Networks". In: *ICML'11: Proceedings of the 28th International Conference on Machine Learning*, pp. 561–568.
- Rupasinghaa, A. and S. J. Goetz (2007). "Social and Political Forces as Determinants of Poverty: A Spatial Analysis". In: *The Journal of Socio-Economics* 36.4, pp. 650–671.

- Senior, Andrew W., Lisa M. Brown, A. Hampapur, C. -F. Shu, Y. Zhai, R. S. Feris, Y. Tian, S. Borger, and Christopher R. Carlson (2007). "Video Analytics for Retail". In: *AVSS'07: IEEE International Conference on Advanced Video and Signal based Surveillance*, pp. 423–428.
- Shadbolt, N., K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, H. Wendy, and M. C. Schraefel (2012). "Linked Open Government Data: Lessons from Data.gov.uk". In: *IEEE Intelligent Systems* 27.3, pp. 16–24.
- Sheldon, D. and Thomas G. Dietterich (2011). "Collective Graphical Models". In: *NeurIPS'11: Advances in Neural Information Processing Systems*, pp. 1161–1169.
- Sheldon, D., T. Sun, A. Kumar, and Thomas G. Dietterich (2013). "Approximate Inference in Collective Graphical Models". In: *ICML'13: Proceedings of the 30th International Conference on Machine Learning*, pp. 1004–1012.
- Smith, C. -C., A. Mashhadi, and L. Capra (2014). "Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks". In: *CHI'14: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 511–520.
- Smith, M. T., M. A. Álvarez, and N. D. Lawrence (2018). "Gaussian process regression for binned data". In: *arXiv e-prints*.
- Song, X., Q. Zhang, Y. Sekimoto, and R. Shibasaki (2014). "Prediction of Human Emergency Behavior and Their Mobility Following Large-Scale Disaster". In: *KDD'14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5–14.
- Sturrock, H. J. W., J. M. Cohen, P. Keil, A. J. Tatem, A. L. Menach, N. E. Ntshalintshali, M. S. Hsiang, and Roland D Gosling (2014). "Fine-scale malaria risk mapping from routine aggregated case data". In: *Malaria Journal* 13, p. 421.
- Sun, T., D. Sheldon, and A. Kumar (2015). "Message Passing for Collective Graphical Models". In: *ICML'15: Proceedings of the 32nd International Conference on Machine Learning*, pp. 853–861.
- Tanaka, Y., T. Kurashima, Y. Fujiwara, T. Iwata, and H. Sawada (2016). "Inferring Latent Triggers of Purchases with Consideration of Social Effects and Media Advertisements". In: *WSDM'16: Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pp. 543–552.
- Tanaka, Y., T. Iwata, T. Kurashima, H. Toda, and N. Ueda (2018). "Estimating Latent People Flow without Tracking Individuals". In: *IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3556–3563.

- Tanaka, Y., T. Iwata, T. Tanaka, T. Kurashima, M. Okawa, and H. Toda (2019a). "Refining Coarse-grained Spatial Data Using Auxiliary Spatial Data Sets with Various Granularities". In: *AAAI'19: Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 5091–5100.
- Tanaka, Y., T. Tanaka, T. Iwata, T. Kurashima, M. Okawa, Y. Akagi, and H. Toda (2019b). "Spatially Aggregated Gaussian Processes with Multivariate Areal Outputs". In: *NeurIPS'19: Advances in Neural Information Processing Systems*, pp. 3000–3010.
- Taylor, B. M., R. Andrade-Pacheco, and H. J. W. Sturrock (2018). "Continuous inference for aggregated point process data". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, p. 12347.
- Teh, Y. W., M. Seeger, and M. I. Jordan (2005). "Semiparametric latent factor models". In: *AISTATS'05: Proceedings of the 10th International Conference on Artificial Intelligence and Statistics*, pp. 333–340.
- Titsias, M. (2009). "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In: *AISTATS'09: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 567–574.
- Vandal, T., E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly (2017). "DeepSD: Generating High Resolution Climate Change Projections Through Single Image Super-resolution". In: *KDD'17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1663–1672.
- (2018). "Generating High Resolution Climate Change Projections through Single Image Super-Resolution: An Abridged Version". In: *IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 5389–5393.
- Wang, H., D. Kifer, C. Graif, and Z. Li (2016). "Crime Rate Inference with Big Data". In: *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 635–644.
- Wang, L., Y. Zheng, X. Xie, and W. Y. Ma (2008). "A flexible spatio-temporal indexing scheme for large-scale GPS track retrieval". In: *MDM'08: Proceedings of the 9th International Conference on Mobile Data Management*, pp. 1–8.
- Wilby, R. L., S. P. Zorita, E. Timbal, B. Whetton, and L. O. Mearns (2004). *Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods*. URL: <http://www.narccap.ucar.edu/doc/tgica-guidance-2004.pdf>.
- Wilson, K. and J. Wakefield (2018). "Pointless spatial modeling". In: *Biostatistics*.

- Wotling, G., C. Bouvier, J. Danloux, and J. -M. Fritsch (2000). "Regionalization of Extreme Precipitation Distribution Using the Principal Components of the Topographical Environment". In: *Journal of Hydrology* 233.1-4, pp. 86–101.
- Xavier, A., M. B. C. Freitas, M. D. S. Rosário, and R. Frago (2016). "Disaggregating statistical data at the field level: An entropy approach". In: *Spatial Statistics* 23, pp. 91–103.
- Xu, F., Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin (2017). "Trajectory Recovery from Ash: User Privacy is not preserved in aggregated mobility data". In: *WWW'17: Proceedings of the 26th International Conference on World Wide Web*, pp. 1241–1250.
- Xu, J. (2017). "Multi-task Learning and Its Application to Geospatio-Temporal Data". In: *ProQuest Dissertations Publishing*. DOI: 10602530.
- Xu, J., X. Liu, T. Wilson, P. N. Tan, P. Hatami, and L. Luo (2018). "MUSCAT: Multi-Scale Spatio-Temporal Learning with Application to Climate Modeling". In: *IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2912–2918.
- Yao, H., F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li (2018). "Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction". In: *AAAI'18: Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 2588–2595.
- Yao, H., X. Tang, H. Wei, G. Zheng, and Z. Li (2019). "Revisiting Spatial-Temporal Similarity: A Deep Learning Framework for Traffic Prediction". In: *AAAI'19: Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 5668–5675.
- Yousefi, F., M. T. Smith, and M. A. Álvarez (2019). "Multi-task Learning for Aggregated Data using Gaussian Processes". In: *NeurIPS'19: Advances in Neural Information Processing Systems*, pp. 15050–15060.
- Yuan, J., Y. Zheng, and X. Xie (2012). "Discovering Regions of Different Functions in a City Using Human Mobility and POIs". In: *KDD'12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 186–194.
- Zhang, J., Y. Zheng, and D. Qi (2017). "Deep Spatio-Temporal Residual Networks for City-wide Crowd Flows Prediction". In: *AAAI'17: Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 1655–1661.
- Zhang, J., Y. Zheng, J. Sun, and D. Qi (2019). "Flow Prediction in Spatio-Temporal Networks Based on Multitask Deep Learning". In: *IEEE Transactions on Knowledge and Data Engineering*. ISSN: 1041-4347. DOI: 10.1109/TKDE.2019.2891537.
- Zheng, Y., F. Liu, and H. P. Hsieh (2013). "U-Air: When Urban Air Quality Inference Meets Big Data". In: *KDD'13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1436–1444.

- Zheng, Y., X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li (2015). "Forecasting Fine-grained Air Quality Based on Big Data". In: *KDD'15: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2267–2276.
- Zhuang, C., N. J. Yuan, R. Song, X. Xie, and Q. Ma (2017). "Understanding People Lifestyles: Construction of Urban Movement Knowledge Graph from GPS Trajectory". In: *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3616–3623.
- Álvarez, M. A., L. Rosasco, and N. D. Lawrence (2012). "Kernels for Vector-Valued Functions: A Review". In: *Foundations and Trends® in Machine Learning* 4.3, pp. 195–266. ISSN: 1935-8237. DOI: 10.1561/22000000036. URL: <http://dx.doi.org/10.1561/22000000036>.

List of Publications

Journal papers

1. Maya Okawa, Yusuke Tanaka, Takeshi Kurashima, Hiroyuki Toda, and Tomohiro Yamada, Marked Temporal Point Processes for Trip Demand Prediction in Bike Sharing Systems, In *IEICE Transactions on Information and Systems*, Vol. E102-D, No. 9, pp. 1635 – 1643, Sep. 2019.
2. 河井悠佑, 田中佑典, 戸田浩之, 石川佳治, 少量の移動履歴を考慮した大規模な集計データからの人流推定, 日本データベース学会和文論文誌, Vol.17-J, No. 7, Mar. 2019.
3. 田中佑典, 倉島健, 藤原靖宏, 岩田具治, 澤田宏, 社会的影響とメディア広告の影響を考慮した購買要因推定, 情報処理学会論文誌, Vol. 58, No. 2, pp. 580 – 593, Feb. 2017.
4. 田中佑典, 上田修功, 田中利幸, クラス固有の特徴選択に基づくベイズ識別器, 電子情報通信学会論文誌 D, Vol. J96-D, No. 11, pp. 2755 – 2764, Nov. 2013.

International conferences (refereed)

1. Yasunori Akagi, Takuya Nishimura, Yusuke Tanaka, Takeshi Kurashima, Hiroyuki Toda, Exact and Efficient Inference for Collective Flow Diffusion Model via Minimum Convex Cost Flow Algorithm, In *AAAI'20: Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020 (to appear).
2. Yusuke Tanaka, Toshiyuki Tanaka, Tomoharu Iwata, Takeshi Kurashima, Maya Okawa, Yasunori Akagi, and Hiroyuki Toda, Spatially Aggregated Gaussian Processes with Multivariate Areal Outputs, In *NeurIPS'19: Advances in Neural Information Processing Systems*, pp. 3000 – 3010, 2019.
3. Maya Okawa, Tomoharu Iwata, Takeshi Kurashima, Yusuke Tanaka, Hiroyuki Toda, and Naonori Ueda, Deep Mixture Point Processes: Spatio-temporal Event Prediction with Rich Contextual Information, In *KDD'19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 373 – 383, 2019.

4. Yusuke Tanaka, Tomoharu Iwata, Toshiyuki Tanaka, Takeshi Kurashima, Maya Okawa, and Hiroyuki Toda, Refining Coarse-grained Spatial Data using Auxiliary Spatial Data Sets with Various Granularities, In *AAAI'19: Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 5091 – 5100, 2019.
5. Yusuke Tanaka, Tomoharu Iwata, Takeshi Kurashima, Hiroyuki Toda, and Naonori Ueda, Estimating Latent People Flow without Tracking Individuals, In *IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3556 – 3563, 2018.
6. Yusuke Tanaka, Takeshi Kurashima, Yasuhiro Fujiwara, Tomoharu Iwata, and Hiroshi Sawada, Inferring Latent Triggers of Purchases with Consideration of Social Effects and Media Advertisements, In *WSDM'16: Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pp. 543 – 552, 2016.
7. Naonori Ueda, Yusuke Tanaka, and Akinori Fujino, Robust Naive Bayes Combination of Multiple Classifications, In *FMI'13: Proceedings of the Forum Math-For-Industry*, pp. 141 – 155, 2013.
8. Takuya Kitamura, Shigeo Abe, and Yusuke Tanaka, Multiple Nonlinear Subspace Methods Using Subspace-based Support Vector Machines, In *ICMLA'11: Proceedings of the 12th International Conference on Machine Learning and Applications*, pp. 358 – 363, 2011.

International workshop (refereed)

1. Yoshiaki Takimoto, Yusuke Tanaka, Takeshi Kurashima, Shuhei Yamamoto, Maya Okawa, and Hiroyuki Toda, Predicting Traffic Accidents with Event Recorder Data, In *PredictGIS'19: Proceedings of the ACM SIGSPATIAL Workshop on Prediction of Human Mobility*, pp. 11 – 14, 2019.

Domestic workshop (refereed)

1. 田中佑典, 倉島健, 藤原靖宏, 岩田具治, 澤田宏, 社会的影響と広告の影響とを考慮した購買行動モデルに基づく購買要因の推定, 第8回Webとデータベースに関するフォーラム (WebDB Forum), pp. 197 – 204, 2015.

Domestic workshops (not refereed)

1. 赤木康紀, 西村拓哉, 田中佑典, 倉島健, 戸田浩之, ネットワークフローを利用したCollective Flow Diffusion Modelの効率的な推定, 第110回人工知能基本問題研究会 (SIG-FPAI), pp. 6 – 11, 2019.

2. 瀧本祥章, 田中佑典, 倉島健, 山本修平, 大川真耶, 戸田浩之, ドライブレコーダデータに基づくヒヤリハット発生予測, 第10回データ工学と情報マネジメントに関するフォーラム (DEIM), 2019.
3. 田中佑典, 岩田具治, 田中利幸, 倉島健, 大川真耶, 戸田浩之, 多様な解像度の補助空間データを用いた空間データの高解像度化, 第21回情報論的学習理論ワークショップ (IBIS), 2018.
4. 大川真耶, 倉島健, 田中佑典, 戸田浩之, バイクシェアシステムにおけるトリップ予測, 第32回人工知能学会全国大会 (JSAI), 2018.
5. 河井 悠佑, 田中 佑典, 戸田 浩之, 石川 佳治, 少数ユーザの移動履歴を考慮した大規模な集計データからの人流推定, 第9回データ工学と情報マネジメントに関するフォーラム (DEIM), 2018.
6. 田中佑典, 戸田浩之, 湯口昌宏, チームスポーツにおける数的優位に至る過程の分析, 第31回人工知能学会全国大会 (JSAI), 2017.
7. 田中佑典, 岩田具治, 倉島健, 戸田浩之, 上田修功, 集計データに基づく人流グラフ推定技術の検討, 第19回情報論的学習理論ワークショップ (IBIS), 2016.
8. 田中佑典, 倉島健, 藤原靖宏, 岩田具治, 澤田宏, 社会環境から受ける影響を考慮した選択行動モデルの検討, 第28回人工知能学会全国大会 (JSAI), 2014.
9. 上田修功, 田中佑典, 中島直樹, メタ学習に基づく加速度センサからの看護師行動識別, マルチメディア, 分散, 協調とモバイルシンポジウム (DI-COMO), pp. 663 – 667, 2013.
10. 田中佑典, 上田修功, 田中利幸, ベイズモデルに基づく判別特徴選択, 第88回数理モデル化と問題解決研究会 (MPS), pp. 1 – 6, 2012.
11. 北村拓也, 藪脇諒佑, 田中佑典, 阿部重夫, 多重非線形部分空間法に基づくサポートベクトルマシンの開発, 知能システムシンポジウム, pp. 259 – 264, 2011.