國立臺灣大學生物資源暨農學院生物環境系統工程學系

博士論文

Department of Bioenvironmental Systems Engineering
College of Bioresources and Agriculture
National Taiwan University

Doctoral Dissertation

新資料時代下以風險評估為觀點的環境資料分析應用

Applications of Environmental Data Analysis From the Perspective on Risk Assessment in the New Data Era

李杰翰
Lee, Chieh-Han

指導教授：余化龍博士
Advisor: Yu, Hwa-Lung, Ph.D.

中華民國 108 年 12 月
December, 2019

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

國立臺灣大學博士學位論文
口試委員會審定書

新資料時代下以風險評估為觀點的環境資料分析應用

Applications of Environmental Data Analysis From the Perspective on Risk Assessment in the New Data Era

本論文係李杰翰君 (D03622009) 在國立臺灣大學生物環境系統工程學系完成之博士學位論文，於民國 108 年 12 月 18 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

Jacqueline MacDonald Gibson

所　長：

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

誌謝

獻給我深愛的家人、朋友與曾經幫助過我的人。

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

Acknowledgements

Dedicated to my beloved family, friends, and whoever helped me.

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

摘要

近年來，在電腦科學與資訊科技領域，無論是在硬體或演算法上取得的巨幅進展，大數據與機器學習皆成為了現今最熱門的兩個詞彙，也使得其他不同領域皆想藉由這股新的科技力量在應用上有所突破。也因為這股熱潮，資料成為了最重要的資產之一；資料分析方法成為了不可或缺的技術；資料科學家成為了人力市場中的熱門職缺。相同地，環境科學領域也極力嘗試結合此新型科技，來找出嶄新的應用方式。然而，由於環境資訊與民眾風險感知是緊緊相扣的，因此，在現今的這股資料浪潮中，從風險分析的觀點，在環境資訊的應用上有許多重要且需被關注的議題。

在本論文中，嘗試由風險評估的觀點出發，探討在現今的的新資料時代下，環境資料分析在應用上其可能性及衍生的重要課題。新資料時代下的多樣應用，加速了政府在開放資料上的進展，然而，環境擁有屬於公共財的特性，環境資料的蒐集與揭露主要掌握在政府部門手中。民眾對於環境資訊知的權利，往往與政府部門形成了對立關係。另外，在政府與民眾對於新科技在環境領域應用上的不熟悉，進而產生環境風險認知上的歧見。在這其中，環境資料科學家藉由其專業的科學知識與能力，在政府部門和民眾之間，形成一交互三角關係。此三角關係中，為了因應新資料時代的發展，每一個角色對於其餘兩個角色皆為不同的利益關係者。

本論文利用三個實際應用案例，作為闡述本論文所提出在新資料時代下，臺灣環境資料分析可能的未來發展方向以及問題所在。首先，本研究在開放政府與開放資料架構下，建立一南臺灣登革熱預警系統，

doi:10.6342/NTU202000011

經由過往難以取得的登革熱發病資料結合氣象因子，提供政府部門在登革熱防治上的預先部署依據，以及民眾對於自身所處環境的登革熱風險認知。第二，本研究利用建立特定商用物聯網空氣感測器的校正模型，經由比較不同可信度監測資料，了解環境數據除數字本身之外，數據的不確定性與民眾風險感知之間的關係，需要謹慎的對待。最後，本研究利用發展具備高效能的資料融合架構，整合確定性與不確定性資料，凸顯在大量含有不確定性的環境資料之下，如何以資料融合方式，達到正確的風險溝通結果。

本論文以風險評估的觀點，檢視現今在這個以資料引領的時代中，環境科學結合資料分析方法在政府、社會與科學三方中所扮演的角色，以及對於環境保護助益的可能性。希望此論文能夠給予未來環境資料分析在風險管理中的一個初步方向。

關鍵字： 環境資料分析，風險分析，預警系統，資料校正，資料融合

doi:10.6342/NTU202000011

Abstract

In recent years, the world has made tremendous progress in computer science and information technology. Either computer hardware development or algorithms evolution lead Big Data and Machine Learning become two most popular words nowadays. Other applied fields also have seen great opportunities on using these emerging technologies to make a breakthrough. Because of this global trend, data has become one of the most valuable asset; data analysis methods have become the essential techniques; data scientists have become the most favored job in human resources market. Likewise, environmental science attempts to apply the new technology and finds innovations. However, environmental information is strongly associated with public risk perception. Hence, there are many important issues from the perspective on risk assessment need to be concerned while surfing on this new data wave.

The dissertation aims to explore application potentials of environmental data analysis and its related issues from the aspect of risk assessment today. The new data era has accelerated the progress of open governmental data. Environmental information is considered as public asset. However, government agencies mostly have authorization of environmental information in collection and reveal. Public's environmental information right-to-know often stands on the opposite side of government agencies. In addition, the reason for the controversy between government agencies and public is unfamiliar with the new technology. Besides, environmental scientists with professional knowledge and expertise forms the interaction triangle with the other charac-

doi:10.6342/NTU202000011

ters that governments, public, and scientists are stakeholders to each other.

This dissertation illustrates the future possibility and problems for Taiwan's environmental data analysis in the new data era by three applications. First of all, under Open Data and Open Government framework, the study constructed an early warning system of dengue fever in southern Taiwan through combining incidences with meteorological factors. The results could provide the disease prevention and control for government agencies and provoke public risk awareness from the disease. Secondly, the study built a calibration model for particular commercial low-cost air quality sensors. By assessing the reliability of measurements, to have understanding that except for the numbers on devices, the relationship of measurement uncertainty and risk perception should be taken into consideration seriously. Lastly, the study developed a high performance data fusion framework that integrated certain and uncertain data to highlight the achievement for proper risk communication with large amount of uncertain environmental information.

The dissertation stands at the perspective of risk analysis to inspect what kind of role that environmental data sciences play in the relationship triangle. In conclusion, the dissertation seeks to open the way for environmental data analysis which is associated with risk management, in further, possible contributions to environmental protection.

Keywords: Environmental data analysis, Risk analysis, Early warning system, Data calibration, Data fusion

doi:10.6342/NTU202000011

Contents

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

List of Figures

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

List of Tables

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

Chapter 1

Personal Thoughts and Experiences on Environmental Analytics

”天色漸漸光，咱就大聲來唱著歌。
一直到希望的光線，照著島嶼每一
個人。”

滅火器 **Fire Ex.** 《島嶼天光》

”Why do I pick *Environmental data analysis* as the theme of my dissertation?” I am asking myself when the journey has come to an end. I was fascinated by the beauty of the Earth when I was still a young child. The beauty has motivated me to dedicate myself to protect this planet since then. At the beginning, I imagined that I would be a more engineer-like person who possibly build some equipments to improve energy efficiency or keep the environment clean. Along with the evolution of computer power and data analytics, after several years, my research interests has slightly been changed. My academic life today is full of data, modelings, programmings, statistics, and information theory etc., and my role is transformed into that is called as *Data Scientist*. Although the transformation has changed skills and tools I owned than what I expected before, my life goal is still the same to be achieved, ”*working towards a sustainable future*”.

Fortunately, I am participating in a world that data is viewed as one of the most valuable resources. This situation gave me great opportunities to solve environmental problems

doi:10.6342/NTU202000011

through data analysis. However, some critical thoughts were generated while I was analyzing the environmental data including questions, observations, suggestions and visions from my personal experiences on environmental data sciences. I would like to express these topics through my dissertation and open for discussion. Following sections are my organized thoughts and idea, and hope the contents will resonate with audiences.

To be clear, these thoughts came from my experiences in Taiwan's environmental studies and most were the participation of government projects.

1.1 The Government Agencies Somehow Get Lost in the Concept of *Open Data*

Taiwan government has been carried out some long-term environmental monitoring projects for decades. Namely, environmental information is not a new type of data, it has been stored and existed for a long time and the public barely paid attention to. Not like western countries, in the past, environmental monitoring data need file applications to be collected in Taiwan. This made the public was difficult to access environmental information freely. For keeping up with *Open Data* trends around the world, the government has worked hard to open environmental information as possible as they can and expects to increase government accountability. However, at first, the government did not know what kinds of data were needed and useful to open to public. A lots of redundant information were released and received complaints from the public. Luckily, with the help of communities and experts, now the government has built a well off open data platform that citizen could easily access to the environmental information after years of working. It also make an impact on conventional environmental management.

In my experience, because the concept of *Open Data* is still a modern and ongoing field, government agencies are difficult to clear see the path of application yet. It definitely need in collaboration with communities or non-governmental organization who have better realization of environmental issues in reality. Although the visibility of environmental information has been raised, the government is confused about the evaluation of applying

doi:10.6342/NTU202000011

security and privacy. *Open Data* between government and public is like a wrestling game. The public has offensive game style that wants all information could be reached; on the other hand, the government plays defensively for security reasons. Furthermore, diverse environmental issues have confounded the governmental effort directions.

One benefit of *Open Data* is to increase government transparency and accountability. But bureaucracy in Taiwan has limited features of environmental open data based on my observations. Government agencies are worried about environmental measurements would be used as political leverage that the public usually doubts on reliability and validity of monitoring. This situation is like a vicious cycle that build on governmental and public concerns each other. It is crucial for government to be balanced against the benefit of *Open Data* and relates to the right to access environmental information.

1.2 The Right to Know: the Conflict Between Government and Public

People are living in the environment and it is a commonsense that everyone should has the right to access and require the environmental information. As a citizen, I would like to know that *"What the quality of air I breathe?"*, *"What the quality of water I drink?"*, *"Does my vegetables grow in the contaminated soil?"*, and *"Is there any infectious disease near my living area?"* etc.. All these information are depends on well-established environmental monitoring systems. As a government, environmental information is not only public service but also represents resource management. Government agencies could determine and execute control strategies and policies by monitoring measurements that these administrative actions could result in increasing government accountability and performance. For example, water is the most valuable resource in the world. In water management, high efficiency of water allocation could be achieved by accurate monitoring and reduce the impact of drought.

Because environmental issues are always connected to human beings or society. Naturally, people have the right to know environmental conditions of places where they live.

doi:10.6342/NTU202000011

There are two famous lawsuits cases about adverse health impacts of groundwater contamination. One occurred in Taiwan and the other was in the U.S.. In the first case, a former U.S. company, Radio Corporation of America (RCA) was found that the company dumped toxic industrial waste in Taoyuan, Taiwan and polluted both the soil and groundwater. Since lack of information about the contamination, employees of RCA were exposed to carcinogens without knowing. The second case in the U.S. alleged contamination of drinking water in the southern California that Pacific Gas and Electric Company (PG&E) discharged waste water and contaminated nearby communities with toxic. Both cases caused huge loss to local people's health, the environment and the companies, and the main reason of these cases is the absence of information about environment at the time.

Environmental information is like a rose has its thorn. Citizen argues about that monitoring equipments and systems were funded from taxation. The information should be considered as public asset and is freely accessible. However, the government is obsessing about environmental information demonstration by public communities and groups, for example, opposition politicians would attack the authority by bad air quality. Data reliability is another critical environmental debate topic. At this point, the government wants to ensure data quality and maintain the accuracy but the public usually thinks about the numbers have been faked. The results of this structure is tended to be a serious contradiction while the role of data is getting more than ever in the future.

1.3 The Temptation of *Big Data* and *Artificial Intelligence* to the Government

In the last decade or two, *Big Data* and *Artificial Intelligence(AI)*, these two words have been widely used in every industry. The tremendous success in information and communication technology in recent years accelerated innovations and the growth of "new economy". To keep up with this global trend, the government has huge investment in development of the new technology. Realistically, if a project is related to either *Big Data* or *AI*, it will get more funding today. It might seem a little absurd since government agencies

doi:10.6342/NTU202000011

believed that entitled projects these two popular words appearing for catching up with the latest technology. Environmental studies are no exception. Do government agencies really know what is the benefit of applying these new technology? Is this kind of technology suitable for environmental researches? To my knowledge, both answers are *NO*.

Government employees in environmental related departments are usually familiar with sample collection works and numerical modelings. They only have basic knowledge in data analytics such as descriptive statistics that is able to describe the basic feature of the data. Therefore, when the new technology has showed up, they only could rely on their imagination of new applications. It would cause huge waste with a smattering of knowledge. In addition, the fancy AI algorithms are not necessarily fit for environmental sciences. Most AI algorithms are associated with training and learning framework which only consider the relationship between input and output variables and ignore the underlying processes. This is a critical issue in environmental studies because of the environment activity is consist of a variety of physical and chemical processes. To understand these underlying processes could be the most important key in environmental sciences. Hence, it is urgent to think about how to find a new way that to apply new technology but not lose the meaningful physical characteristics.

1.4 The Myth of Uncertainty in Environmental Risk Communication

It is a scientists nature that have a sense of limitations and uncertainty in a research no matter is deterministic or stochastic study. Environmental scientists are able to clarify uncertainties including possible sources. Being a scientific researcher, I always thought a research work could be defended successfully in terms of specifically explaining the limitations and uncertainty until I have participated and experienced in some discussions with local communities. I found out that the concept of uncertainty is vague to citizens and they even do not care about what it means. Due to environmental sciences are likely to have close connection to risk assessment, this issue can be referred to be part of risk

doi:10.6342/NTU202000011

communication. Uncertainty information is one of many aspects of risk communication and it is getting more and more important in this new data era. In the era of *Big Data*, data generation process has been changed, e.g., the development of low-cost sensors, it has huge impact on risk awareness and response, especially in environmental measurements.

Today there are a lots of affordable commercial devices on the market and people can easily use the devices with wireless networking to monitor the environmental quality they concerned about anywhere anytime. Therefore, risk awareness and sensitivity are highly dependent on numbers showing on device panels. Without the concept of uncertainty, a number is exactly a number not an interval or probability. Even though fast response and inexpensive of these new devices help significant increment in space-time resolution of environmental monitoring, the existing data quality issue of these low-cost sensors could be a misleading in risk awareness or communication.

In addition to measurement itself, there is a responsibility issue caused by the difference of viewpoints from scientific fields. From the point of view of computer science or information technology(IT), the quantity which measured by sensors is a matter of objective and impartial. ”Let data speak for itself” is often used to describe this situation. Although many experts and organizations have made corrections to this fault notion, I still see lots of IT people shows the irresponsibility to not concern about what information they have published. On the opposite side, environmental sciences emphasize the meaning behind the measurement numbers. Environmental scientists are afraid of reporting a result that can not be explained, therefore, they are very carefully using this new kind of measurements. However, this act would be considered conservative and go in the opposite direction of *Big Data*.

I have attended to a conference in Taiwan about applications of new IoT technologies on environmental monitoring. In the conference, scientists from computer science and environmental science were against each other. The major argument was the reliability problem of new environmental monitoring technology and equipments. In my opinion, clear and effective communications could reduce public worry and stress that take more responsibility for the society. The problem we have to think about is how to become a

doi:10.6342/NTU202000011

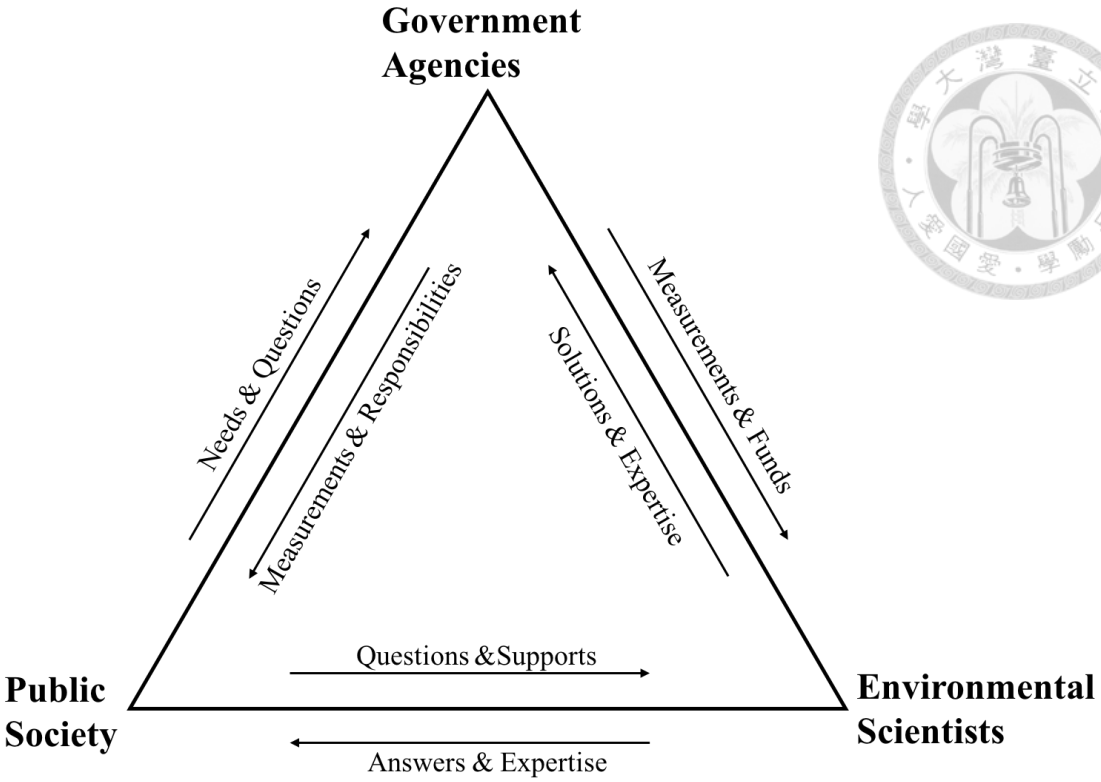


Figure 1.1: The triangle relationship between government agencies, environmental scientists, and public society.

truly neutral environmental information provider. Not only provide the numbers but also its representations or sciences would become a protective umbrella.

1.5 Ideal-Practice Gap in Environmental Sciences with Data Analysis

On my mind, the gap of ideally and practical environmental data analysis is formed by a triangle relationship based upon accountability and responsibility. The triangle relationship is constructed by three characters including government agencies, environmental scientists, and public society. All three characters have played against each other and held their own perspective to environmental information. The triangle relationship as shown in Figure1.1.

Government starts off the purpose of regulation and management that nature is regarded as property. A duty of a government is to maintain a sustainable living environ-

doi:10.6342/NTU202000011

ment which is called environmental governance. Government agencies in order to respond to the public expectation of providing a better living environment, they seek help from environmental expertise by building environmental monitoring infrastructures and funding projects. Government can be as an environmental information provider in this relationship triangle; Citizen has the strongest conscious and willingness to pursue environmental protection. Accessibility of environmental information is the most concerning topic for local environmental related communities and associations. Citizen also need the help from experts for answering and supporting their questions, in further, asking government for enhancement of environmental policies. Although citizen could collect data by themselves nowadays, lack of knowledge make them with restricted sense about environmental information, e.g., they would questions about what factors caused bad air quality. In this relationship, citizen plays a role of information receiver; At last, environmental scientists could be both an information provider and receiver in this triangle. Environmental studies have their special characteristics representations. As a provider, environmental scientists could illustrate issues precisely with their expertise and raise environmental information values in advance. On the other hand, as a receiver, environmental scientists not just collect ”real” data, furthermore, they have social responsibility to gather environmental and social concerns.

This complex triangle structure create the gap between ideal and practice. The three roles are interfering with each other. On many occasions, one’s goal would be held back by the others. Battles inside the triangle even lead to the problem of moral responsibility. As a environmentalist, it seems to me that the biggest difference between environmental sciences and other scientific fields is that almost every research project is funded by government agencies. Due to this fact that the flexibility and energy of environmental researches somehow have been limited. The last section has mentioned that it is complicated to elucidate who is responsible for the reveal of environmental information. Unlike a recommendation system in commercial applications. It is difficult and nearly no freedom of choice when environmental information has been published. The impact of environmental information is very huge to the public and wrong interpretation is even more harmful. At

doi:10.6342/NTU202000011

this present time, no matter is *Big Data* or *Open Data*, it seems like a double-edged sword to environmental science to me. As an environmental scientist, it is worth to think about how to play a role and stay balanced inside this triangle. Environmental scientists not only can make contributions to the society with specific expertise but are responsible for the outcomes from the interpretation of environmental information.

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

Chapter 2

The Critical Aspects of Environmental Data Analysis in the Present

”我們把希望寄託在另一個世界裡，然後才能面對這殘破的生命。有誰又為了別人而犧牲自己，我們又學到了什麼生命的意義。”

那我懂你意思了 **Iguband**《沒有人在乎你在乎的事》

Big Data and *Machine Learning* have been the most popular and commonly used vocabularies in all research topics. From Google Trends which shows the frequency of a given term is entered into the search engine, the worldwide interest of these two vocabularies have showed dramatic increase in the last decade[41]. Based on these two concepts, a bunch of contemporary or new algorithms and technologies have been developed and applied in various fields. And this phenomenon resulted in changing and bringing applied sciences into a new era. Without exception, the environmental sciences also embrace the fashion and provide lots opportunities to the data analysts. However, there are some doubts and questions have emerged while environmental sciences is adapting to these innovative technologies. Due to human beings are living in the environment, environmental research topics mostly related to every kinds of risk assessments, for example, health risk, natural

doi:10.6342/NTU202000011

hazard, and food safety etc. Therefore, it is a critical issue to clarify the possibilities and limitations of the new techs that are applying to the environmental sciences for the public.

In this chapter, first of all, the history of data analysis would be introduced and reviews the evolution of analytics techniques. Next, the concept of modern data and analytics such as large datasets an AIs would be explained to provide the shape of intelligent data analysis. In further, conventional environmental sciences would be described to reveal how the importance of role modern data analytics have played and connected to the influence on risk assessment. Finally, the objectives of this dissertation would be issued.

2.1 From the past towards possible futures of data analysis

While words or phrases that related to ”data analysis”, the first impression of audiences is that to obtain information from some observations. Intuitively, we would like to have meaningful results from the data we have collected. It is like treasure hunters are digging valuable things from a bunch of collections. But, the scientific point of view, data analysis is a much more complex field that involves a variety of professional research topics. The stages of data processing such as inspecting, cleaning, visualizing and modeling data etc. are all considered as parts of data analysis. These techniques are developed to achieve a common goal: discovering and gathering useful information to produce a better outcome in decision-makings or business operations.

Even more than 50 years ago, a statistician, John W. Tukey had defined data analysis as ”procedures for analyzing data techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”[93] He pointed out an as-yet unknown science which was different than neither *mathematics* nor *statistics*. It was still a vague idea at that time. Today, his sense has truly become a certain research field, which we have called as *Data Science* now.

In the beginning, data analysis was particularly applied to the need for business in-

doi:10.6342/NTU202000011

telligence. The main purpose of business intelligence was to transform from raw data into meaningful information which was very crucial for business managements. How to make effective decisions, estimate profit margin, or create successful budget plan, these issues were important for a firm keeping competitive. And data analysis was the key to determine a firm's managing strategies. In business intelligence, data analysis covered data collecting, data cleaning, exploratory data analysis, data modeling, and data communication. It is worth mentioning that there were two main techniques of data modeling: data mining and on-line analytical processing. The difference between data mining and on-line analytical processing is the driving force of algorithms that one is data-driven and the other is model-driven, respectively. Although on-line analytical processing could integrate multiple data processing functions, the advantage of data mining is its ability of future predictions[103]. Data mining is a particular data analysis technique, which is consist of classification, association analysis, cluster analysis, and anomaly detection, focuses on predictive purpose rather than descriptive[8]. Nevertheless, these data analysis techniques have been developed and used for a long time and successfully accomplished the tasks whereas they mostly were utilized within organizations or scientific researches.

Until now, while the users are breaking out of the comfort zone and not limited to the specific field of study, we have seen the tremendous progress in the applications of *Data Science*.

2.2 Big Data and Machine Learning make a different future

After we have made the substantial progress in computer hardware industry, the technologies are no longer just focus on one specific research field. They are pursuing for wider and sophisticated applications to generate extra revenue from the data. The notable example is the rise of social media analytics. This phenomenon not only changes the way people connected and content shared, but give *Data Science* a perfect experimental field in *Big Data*. It is well known that *Big Data* has the four V's, which are *Volume*, *Velocity*,

doi:10.6342/NTU202000011

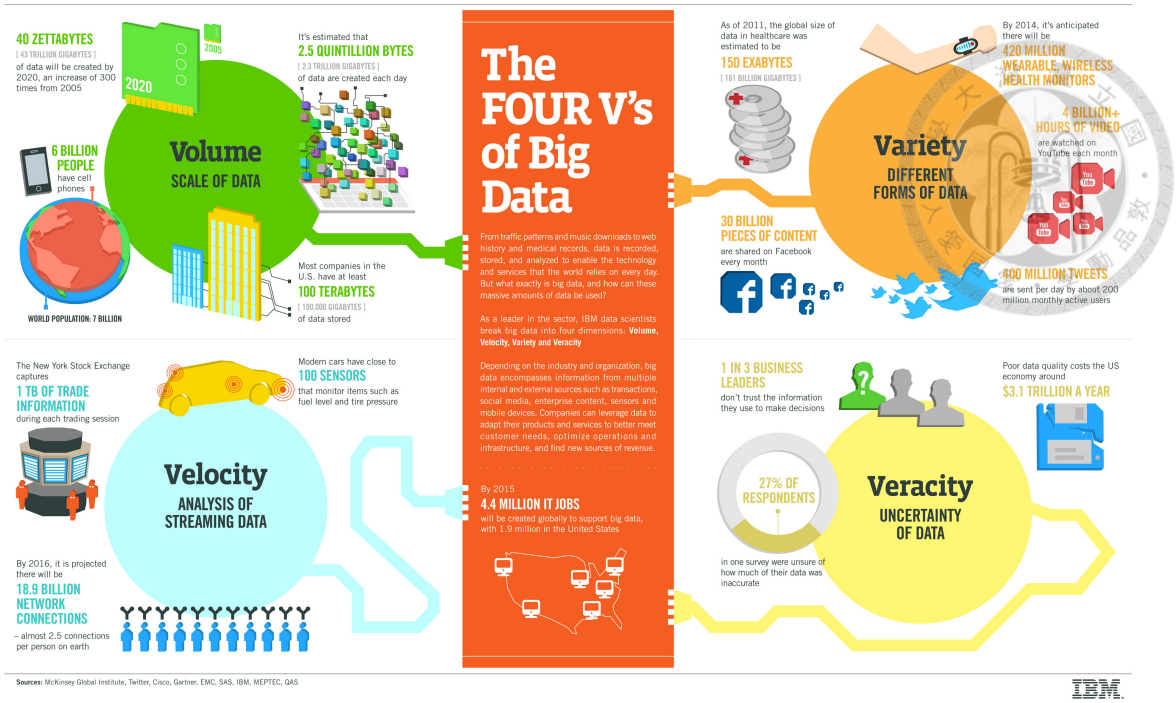


Figure 2.1: The four V's explanation of Big Data by IBM scientists[32].

Variety, and *Veracity* to describe its specificity. IBM Corp. breaks these four V's and has clear explanation by an inforgraphic, as shown in Figure 2.1. Based on these properties, data is not the biggest issue for organizations or scientists. The challenge will become finding the highest-value opportunities, using new technologies and tools to carry out analytics, gaining the benefit, and planning for the future[85]. This data-driven concept even offerers new energetic power to the past developed algorithms such as we can obtain more detailed information by simple data visualization under a large dataset.

AI usually is stated along with *Big Data*, however, AI is not a brand new technology. In general, in contrast to the natural intelligence which means determined by human, AI can be seen as intelligence is demonstrated by machinery agents. In 1950, Alan Turing announced the famed *Turing Test* in computer science[94] and triggered off thoughts waves of using machines to "learning" and "problem solving". People believe that AI could make the society more efficient no matter in manufacturing or decision-making. After more than 60 years, numerous algorithms and tools are related to AI nowadays such as search and mathematical optimization and artificial neural network. *Machine Learning* or *Deep Learning* is the most popular word in terms of AI in recent times. According to

doi:10.6342/NTU202000011

theses modern AI techniques, a astonishing machine, AlphaGo, was created to the world in 2015 and became one of the most breakthrough a year later that the first time a computer beat a professional human Go player[84]. This case illustrates the core and capability of *Big Data* and *Machine Learning* that an efficient machine can predict accurately given a large enough training data.

2.3 Reviews of Environmental data analyzing methods and applications

Environmental data analysis is a specific research field in data analysis that is applied collected data from the environment to a variety of demands. Environmental sciences is consist of many branches of sciences inside including meteorology, ecology, hydrology, physics, chemistry, geology and biology etc.. It is simple to look at environmental data analysis in space and time that could obtain three different environmental analytics: *spatial analysis* and *temporal analysis*, and *spatiotemporal analysis*.

Because of environmental data analysis also attends to discover useful information from the data as other data analytics fields, most of statistical models have been applied. For *temporal analysis*, methods for time series analysis could be used including frequency-domain methods and time-domain methods. Frequency-domain methods are such as spectral analysis and wavelet analysis, e.g., Fourier analysis[87]. Time-domain methods, for example, autoregression(AR), autoregressive integrated moving average(ARIMA) are commonly used in building models[44]. For *spatial analysis*, several regression techniques were developed for environmental analysis, e.g., land use regression[50] and geographical weighted regression[98].

Major environmental data collection is from well-established monitoring stations or sampling by researchers themselves. Data abundance is highly depending on the sampling difficulty and cost. The situation results in both limited spatial and temporal resolution of environmental data. For the reason, several spatial interpolation techniques or geostatistics were arose from the purpose of assessing the environmental information at unknown

doi:10.6342/NTU202000011

locations. The most famous and widely used one is Gaussian process regression as known as Kriging method[60, 23]. Otherwise, since the environment is regarded as a public asset, the most environmental data and projects are founded by government agencies. This circumstance caused the environmental researches are either limited by data abundance or accessibility. Formerly in Taiwan, the accessibility of environment data somehow is relatively restricted by the government. Only could be reached by scientific research purposes or government projects. Hence, environmental data analysis was not received too much attention within citizens and communities.

However, the concept of *Open Data* and *Open Government* has changed the situation at present. In addition, the rapid growth in Internet technologies also raised environmental data analysis to another level.

2.4 Exposure assessment and risk communication in connection with environmental data analysis

Risk analysis contains various phases and will varies among different kinds of risk assessment. Due to the special characteristics of environmental data, the study will focus on two particular phases of risk assessment that are called exposure assessment and risk communication. As mentioned, environmental information is a public asset that citizens should have the right to access the information freely. The importance of knowing the environment conditions is the main idea of exposure assessment. Environmental sciences has played an essential role in exposure assessment because it is an interdisciplinary research field and also needs cross-disciplinary knowledges. Environmental sciences can integrate multiple scientific fields to the study such as meteorology, chemistry, ecology, hydrology, and biology etc. as show in Figure 2.2. Similarly, environmental data analysis is a valuable tool in exposure assessment, like geostatistics which is described in the previous section, that is mainly used to estimate environmental exposures. Estimation of environmental exposures here can have a wide range of applications such as typhoon forecast, air quality forecast, water quality assessment, and food security assessment etc.. These

doi:10.6342/NTU202000011

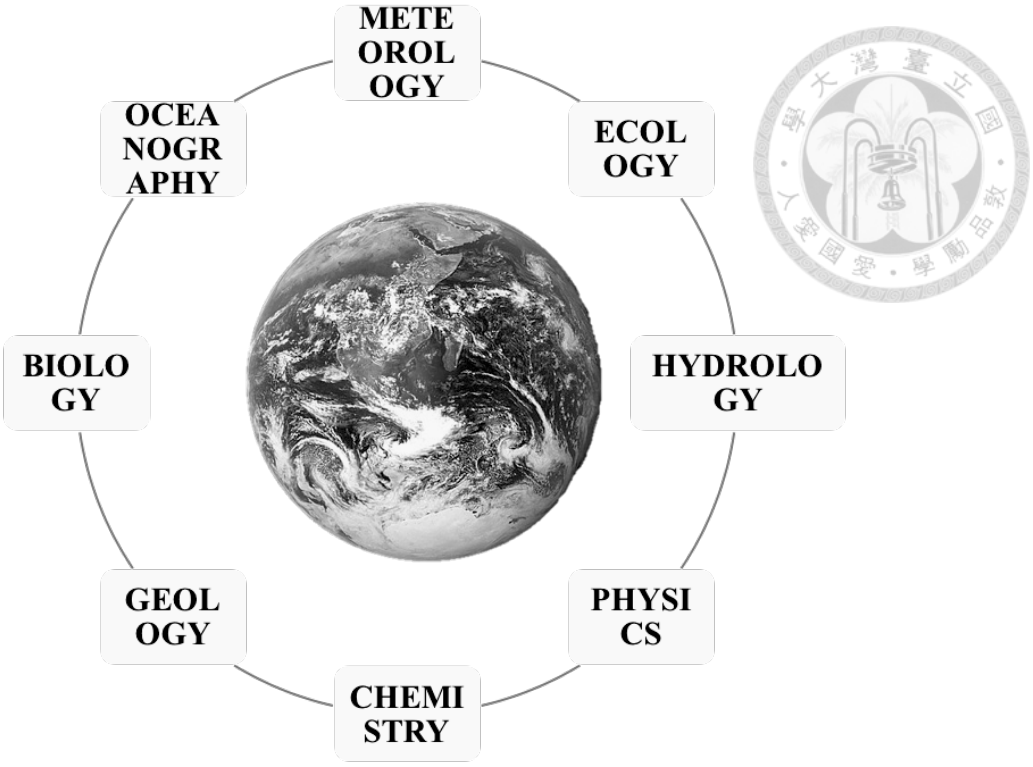


Figure 2.2: Environmental sciences is an interdisciplinary study field.

estimations are highly related to vulnerability, disease burden or adverse health impact, in other words, that is the term we called *risk*.

However, limitations and uncertainty are always existing in sciences and technologies. Although modern sciences have made significant progress comparing to hundreds or thousands years ago, there are still a lots of unknown things to discover. Especially in environmental studies, it is impossible to fully simulate a real world. Each research study could only provide a tiny piece of the whole world even without a 100% correct answer. That is why environmental scientific studies have to illustrate either limitation or uncertainty very clear. The most concern, in reality, is that if the scientific results are used for the purpose of increasing public risk awareness, the uncertainty of results might cause different risk perception to public. This is a crucial issue for environmental scientists because it is impossible to ignore and dismiss all uncertainty of research studies. Hence, we need shift our focus from providing the true answer to communicating under uncertainty.

There is a specific field in risk analysis called *risk communication*. The purpose of risk communication is that scientists or experts intentionally conveying their risk information

doi:10.6342/NTU202000011

to a targeted audience through designated channels[7]. Targeted audiences includes non-experts, the general public, interest groups or regulatory practitioners. After decades in development of risk communication, trust has become a vital aspect of improving risk communication nowadays. It is critical to make persuasive communications about risk information to audiences with the limits of knowledge improving risk awareness and acceptance[64, 12]. Environmental sciences are associated with healtth risks, environmental hazards studies. In other words, environmental data analysis has deeply involved in exposure assessment of risk analysis. The results of environmental data analysis should take the communication techniques into consideration in further.

2.5 A new era of environmental data analysis from the perspective on risk assessment

Today's world is facing a sharing economy also known as collaborative consumption. The tremendous progress made in computing capability in recent years has given huge support to the trend. This sharing trend also delivers a noticeable impact on the growth of *Open Data* and *Open Government*. Nevertheless, these two concepts are often mixed together as "open government data". Harlan Yu and David G. Robinson pointed out the distinction between *Open Data* and *Open Government*. The concept of *Open Government* focuses on increasing a government with more transparency, participation, and collaboration which also means the goal is to enhance civic and political accountability for public. On the other hand, *Open Data* can be referred as any open technology that makes public information more adaptable and provides new aspects of civic life. Although technologies can empower the public data value, it will not resolve the situation for political and managerial accountability for government alone[106].

To my mind, the combination of *Open Data* and *Open Government* provides a perfect opportunity for the environmental data analysis field in Taiwan. There is an idiom in English, "You scratch my back and I'll scratch yours.", ideally describes the relationship between environmental data and Taiwan's government agencies. Previously, environmen-

doi:10.6342/NTU202000011

tal data were only used for the monitoring environmental quality purpose. Due to, at the moment, most of environmental data were inaccessible to the public, a lots doubts of measurements accuracy were arose between citizen and communities. Following the concepts of *Open Data* and *Open Government*, the government can overcome these doubts by sharing the environmental information and the public can contribute with new aspects to the government. In further, to enhance government accountability.

The concept of Internet of Things (IoT) has grown rapidly in the recent decade. There are various IoT applications have been developed now or in near future[42]. One essential key of IoT paradigm is the wireless sensor networks (WSN) that beneath the fast development of Internet. Accordingly, the IoT has brought a new era of environmental monitoring to build a smart city. Low cost and great mobility are the two most attractive features of IoT equipments. *Big Data* and IoT are working well together that *Big Data* analytics can empower IoT to generate valuable information. It seems could be a solution to the dilemma of lack of environmental data. However, while using IoT devices for environmental monitoring, there is a serious issue has emerged and should be carefully considered, *reliability*.

Here we are focusing on uncertainty of IoT device measurement, in other words, data quality is essential to environmental data analytics. Environmentalists have forged to answer about how the planet works through a variety of scientific researches. Meanwhile, their research outcomes have made a great contribution to the society that seeks for risk avoidance or risk reduction. Environmental sciences have been played an important role in whether health risk assessment or natural hazard assessment. Data quality issues have huge impact on risk assessment that uncertain measurements could cause public risk awareness to be misleading. It not only affects social aspects but have a great influence on policymaking and management in further. Hence, *Veracity* of *Big Data* has invoked cautious on use of new type of environmental data and the importance of risk communication to the public.

There were many innovative applications have shown the capability of technology in the new era. Google Flu Trends is one of the most impressive example that discovered

doi:10.6342/NTU202000011

a whole world of valuable *Big Data*. Google scientists announced that they could predict accurately flu prevalence two weeks earlier based on flu tracking information from the Centers for Disease Control and Prevention and flu-related information on Google search engine[39]. However, it failed on flu prevalence estimation few years later and other researchers deconstructed the failure[62]. This case shows *Big Data* still display great potential on environmental analysis but some keys are missing. Mostly, data-driven modeling and techniques are searching for statistical relationship like correlation and dependency. For instance, a huge amount of *Machine Learning* algorithms are considered as black box models which means the models determine the relationship between input and output variables without thinking about underlying physical processes. This framework caused hard to be convinced in risk communication. For environmental scientists, decision makers, and government agencies, it is questionable and difficult to realize a natural system just based upon statistical relationships because of ”correlation is not causality”.

In my opinion, causality is the breaking point of environmental data analysis in the era of *Big Data*. Previously, economists have proposed that causality could be learned and then inference[6]. Besides, another straightforward way in environmental sciences to address this problem is through integration of statistical and physical models called data assimilation or data fusion techniques. Within this framework, physical laws and other governor equations of numerical models become constraints in statistical modeling. On the other hand, features or patterns found by statistical models could improve the performance of parameters estimation as general knowledges in numerical models. Hence, these approaches could be expected to create a win-win situation and produce more persuasive research results.

doi:10.6342/NTU202000011

Chapter 3

Objectives of the Dissertation

”你是追尋玫瑰的獨角獸，強忍淚水體會成長，時不時的替你擔憂。露水的營養，朝陽的變化，迷霧的溫度，山嵐和彩霞。”

告五人 **Accusefive**《獨角獸》

From Chapter 1 (*my personal experiences in environmental data analysis*) and Chapter 2 (*my thinking and vision as an environmental data scientist*), in my mind, environmental sciences in this new data era are not just scientific but more close to philosophical subjects. As mentioned, due to the specificity of the environment-related issues, there are three important characters have participated and formed a triangle relationship including governments, scientists, and citizens, that all characters interact with each other. There are different positive and negative effects between their relationships. Moreover, while *Big Data* and *AI* have revolutionized environmental data analytics, everyone wants to seize the opportunity to keep up with these hot trends. Thus, the measurements of environmental monitoring have become the most valuable resources for environmental data analysis. Environmental data analysis has become a complex topic. In the next three Chapters, this dissertation will use three applications of environmental data analysis to answer three questions as follows.

Along with *Open data* and *Open government* trends in Taiwan, environmental infor-

doi:10.6342/NTU202000011

mation which had seen as governmental property that have been largely released recently. This result initiates the first question: ”How do we use these environmental data based on the *open* concept?”.

- In Chapter 4, a spatiotempoal early warning system of Dengue fever will be introduced to illustrate the benefit of environmental information opening and show the great predictability that uses the combination of epidemiology and meteorology to forecast the incidences diffusion.

After the rapidly growing in IoT industry, now people can easily used low-cost sensors to monitor the environment. There are numerous commercial devices on the market can choose and have generated a large amount of environmental measurements. Here comes the second question: ”How to quantify the reliability of low-cost sensors measurements?”.

- In Chapter 5, a spatiotemporal calibration approach for low-cost $PM_{2.5}$ sensors will be introduced to point out the existence of low-cost sensor measurement bias and discuss the bias affects public risk perception from air pollution.

Due to the evolution of environmental monitoring, certain and uncertain measurements are existing at the same time. Under the consideration of different kinds of uncertainty and value representation in these two data types, there is the third question: ”How to integrate multiple data sources with different uncertainty?”.

- In Chapter 6, a high performance data fusion technique will be introduced to exhibit the integration of regulatory stations and low-cost sensors $PM_{2.5}$ measurements and focus on the visualization of air pollution mapping which is strongly related to issues of risk communication.

doi:10.6342/NTU202000011

Chapter 4

A Spatiotemporal Dengue Fever Early Warning Model Accounting for Nonlinear Associations with Hydrological Factors: a Bayesian Maximum Entropy Approach

(*Published in Stochastic Environmental Research and Risk Assessment, 2016[108])*

"But I say you'll see, I'd make you see. Every detail of this damned life. I say you'll see, I'd make you see. You'll see。"

Tizzy Bac《You'll See》

In the last decade, governments around the world have made great efforts to create Open Government Data (OGD). Either to improve administrative transparency or enhance citizen participation. It is an unstoppable trend that governments provide their data assets and empower the public to boost government expenditure data through cross domain knowledge integration, innovative algorithms and enormous computing power into this

doi:10.6342/NTU202000011

new epoch. As a result, governments would benefit from new features of applications and give better decision-making quality; the public and personal could have better understand and monitor the government operations.

In spite of finance, environmental monitoring is the most closely governmental data related to citizen. Accompany with environmental awareness grew rapidly in the second half of the 21th century, people have paid more attention to the quality of where they are living with such as water quality, air pollution, infectious disease and food safety. Environmental open data shows one of the greatest potential opportunities for advanced data science applications. As mentioned, environmental awareness can be also recognized as risk awareness of livings. Environmental data usually is used for exposure assessment in risk assessment which to evaluate and quantify hazard intensity. Scientists could take advantage of a variety of environmental variables and find out the relationships between them. In addition to the large data volume, one special property of environmental monitoring is to support environmental information in near real-time observations. With real-time data, the analytics can easily transition from historical analyzing to forecasting modeling.

In epidemiology, for the purpose of diseases prevention and control, researchers are always concerned about predicting the pattern of infectious diseases. It is an apparent application that is performed with both historical and real-time data. Typically, the public was difficult to obtain incidence datasets in Taiwan. However, in recent times, Taiwan government's open data platform has changed the circumstance that citizen can easily access the database now. Here we will use dengue fever as an example of the potential environmental open data analysis. In this chapter, it will introduce an early warning system of dengue fever that integrates a distributed lagged nonlinear model with geostatiscal approach based on open datasets to illustrate the importance and influence of open government data.

doi:10.6342/NTU202000011

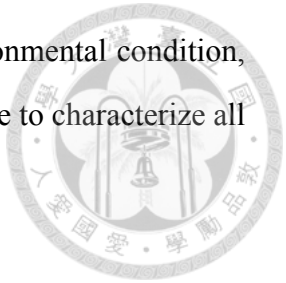
4.1 The relationship between dengue fever and meteorology

Dengue fever (DF) has been recognized as one of the most important international epidemic concern in recent decades. It is a mosquito-borne viral infection that is found in tropical and sub-tropical climates regions around the world. DF can lead to headache, nausea, vomiting, and joint pains, and its severe form, i.e. dengue hemorrhagic fever, is a leading cause of death and serious illness among children in some Asian countries[99]. Dengue incidences have grown dramatically in the past years. Over 40% of the world's population is now at risk from dengue and about 50–100 million people are infected by dengue worldwide annually. Dengue viruses are primarily transmitted through the bite of infected mosquito vector Aedes aegypti in urban and suburban areas. At present, there is no vaccine and specific treatment for dengue fever. The most effective prevention and control are to take anti-mosquito measures by environmental management (e.g., removing artificial habitats and emptying water containers routinely). Thus, establishing a DF early warning system (EWS) to predict the DF occurrences across space and time with a sufficient lead-time is urgently needed for the preventive or control purposes.

In general, previous studies have found significant relationship between the spatial or spatio-temporal epidemic spreading patterns of DF and hydrological-related factors[11, 102, 109, 16, 15]. Among them, temperature plays a substantial role in the development of mosquitoes. It's not only influence on the proportion of infective population but survival of the vector including hatching rate, size of larvae or mortality[104, 105]. Nonlinear associations can exist between temperature and vector development in that temperature increase can be positively or negatively associated with the development of mosquitoes at different temperature ranges[73, 76]. Precipitation is the key of eggs and larval development[65, 77, 86]. Although heavy rainfall can destroy the habitats of mosquitoes, precipitation is supplying the vector with suitable breeding environments like pools, water-filled container and other aquatic habitats. However, the transmission of the infectious diseases across space and time is a complex interactive process including participation of human

doi:10.6342/NTU202000011

hosts, the virus serotype, and environmental factors. An epidemic disease outbreak must consist of three elements, human host, existence of virus, and environmental condition, which can change across space and time; nevertheless, it is a challenge to characterize all factors.



4.2 Early warning system modeling for dengue fever incidences

For the purposes of disaster preparedness and emergence response, EWS development has become an emerging practice for infectious diseases control in recent decade. A disease EWS can provide crucial information for public health enabling to take actions on preventing and mitigating the impact of potential diseases outbreak. Development of disease EWS has been become an important strategy for the disease prevention and mitigation by investigating the statistical associations between environmental conditions and disease incidences. In the case of DF disease, numerous studies have applied a variety of statistical models, e.g., generalized linear model, Poisson regression and mixed model[22, 67, 68, 109, 46], to investigate the relationships between the time series of hydrological factors and DF incidences for the purposes of disease prediction or EWS development, e.g. time lag effects[47, 48, 101, 27]. The identification of this temporal relationship, i.e., time lag effects, is challenging due to the high autocorrelation in the time series of both hydrological and disease data. Recent studies have applied the distributed lag model (DLM) and its variant, i.e. distributed lag nonlinear model (DLNM), to identify the lagged relationships and mitigate the multi-collinear issues[111, 80, 89, 27]. In addition to the understanding the statistical association between hydrological related factors and DF incidences, disease diffusion also plays an important role in spatiotemporal distribution of disease incidences. The disease diffusion is primarily dominated by population dynamics, e.g. the population movement, and the interactions between the populations of human and vectors. Many epidemiological models have been proposed to characterize the diffusion of infectious diseases, including gravity, point process, spatial micro-simulation, network-

doi:10.6342/NTU202000011

based models, and susceptible-infected-recovered model[34, 9, 79, 78, 33, 29, 46]; however, most of these models require the estimation of extensive parameters which are commonly unknown due to the limited knowledge of the disease spread. Geostatistical methods, e.g., Kriging and Bayesian maximum entropy (BME) methods, have been shown to be a proper surrogate approach for the modeling of system dynamics with high complexity in parameters, and was used to characterize the space–time patterns of disease diffusion[82, 81, 109, 5, 53]. Despite of all these approaches in disease modeling, the studies integrating both disease dynamics and associations with external forcing are still limited[72]. Yu et al.[109] considered both disease dynamics and external forcing for spatiotemporal modeling with assumption of the linearity between the hydrological changes and logarithm of DF risks, in which multicollinearity among hydro-climatic predictors were shown.

In order to develop an EWS for the 1-week-ahead prediction of spatiotemporal disease spread of dengue fever, this study proposed an integration of DLNM and an epistemic-based geostatistical approach, i.e. BME method, to account for both disease dynamics and hydrological influence to the spatiotemporal distribution of DF incidences, and to mitigate the multicollinearity issues among the space–time dataset. Among them, DLNM is used to reveal the space–time lagged relationships between hydrological and dengue fever processes, and BME is used to characterize the spatiotemporal DF diffusion with considering the data uncertainties. We applied our approach to assess the dengue fever epidemics across space and time in southern Taiwan for the period of 1998–2012.

4.3 Dengue fever in southern Taiwan

This study investigated the DF incidence cases in the major epicenter of dengue fever in Taiwan, i.e. tropical and south part of Taiwan including Tainan City, Kaohsiung City, and Pingtung County. The annual incidence in the study area is very high; the number of DF cases in southern Taiwan constituted over 94 % of the total number of cases in entire Taiwan during the past decades. In this study, the DF observations of southern Taiwan are based on surveillance data obtained from the Taiwan Center of Disease Control during

doi:10.6342/NTU202000011

the period of 1998–2012. The DF dataset consists of confirmed incidences (number of reported cases) sorted into the temporal and spatial scales of week and district (an administrative unit), respectively. According to the standard surveillance procedure in Taiwan, every suspected DF case should be reported to medical laboratories for the confirmation within 24 h from its diagnosis by clinics or hospitals. In general, the reported DF cases should take about 1–3 days to be confirmed in our study area. About 63 % reported cases were confirmed to be positive. On the other hand, surveying showed that only 83 % suspected cases were reported[52]. Figure 4.1 shows the averaged annual DF cases across the study area. Figure 4.2 shows the weekly-based temporal variation of total recorded of DF cases during the study period as well as those of associated hydrological factors, i.e. the averaged temperature, the maximum temperature, the minimum temperature, the total rainfall, the maximum 24-h rainfall, and the maximum 1-h rainfall. These hydrological data were collected from the monitoring stations of Taiwan Central Weather Bureau. An inverse distance weighting method was used to estimate weekly hydrological measurements for each district, i.e., estimations at the centroids of the districts[35], due to the comprehensive spatial coverage of the monitoring stations, shown in Figure 4.1.

4.4 Spatiotemporal DF prediction

4.4.1 BME method

BME method is a geostatistical approach based upon an epistemic framework. It considers the space–time distribution of DF incidences as a spatiotemporal random field (S/TRF), in which $Z(p)$, $p = (s, t)$, represents DF cases in the study area, where the vector s and t denotes the location (township) and the observed time of dengue incidences by calendar week, respectively. The BME approach distinguishes two major knowledge bases (KB) for the spatiotemporal modeling:

(a) Core or general knowledge base (G-KB) that can include all knowledge bases of natural characteristics, which can be scientific laws, empirical relationships, and theoretical space–time dependence models.

doi:10.6342/NTU202000011

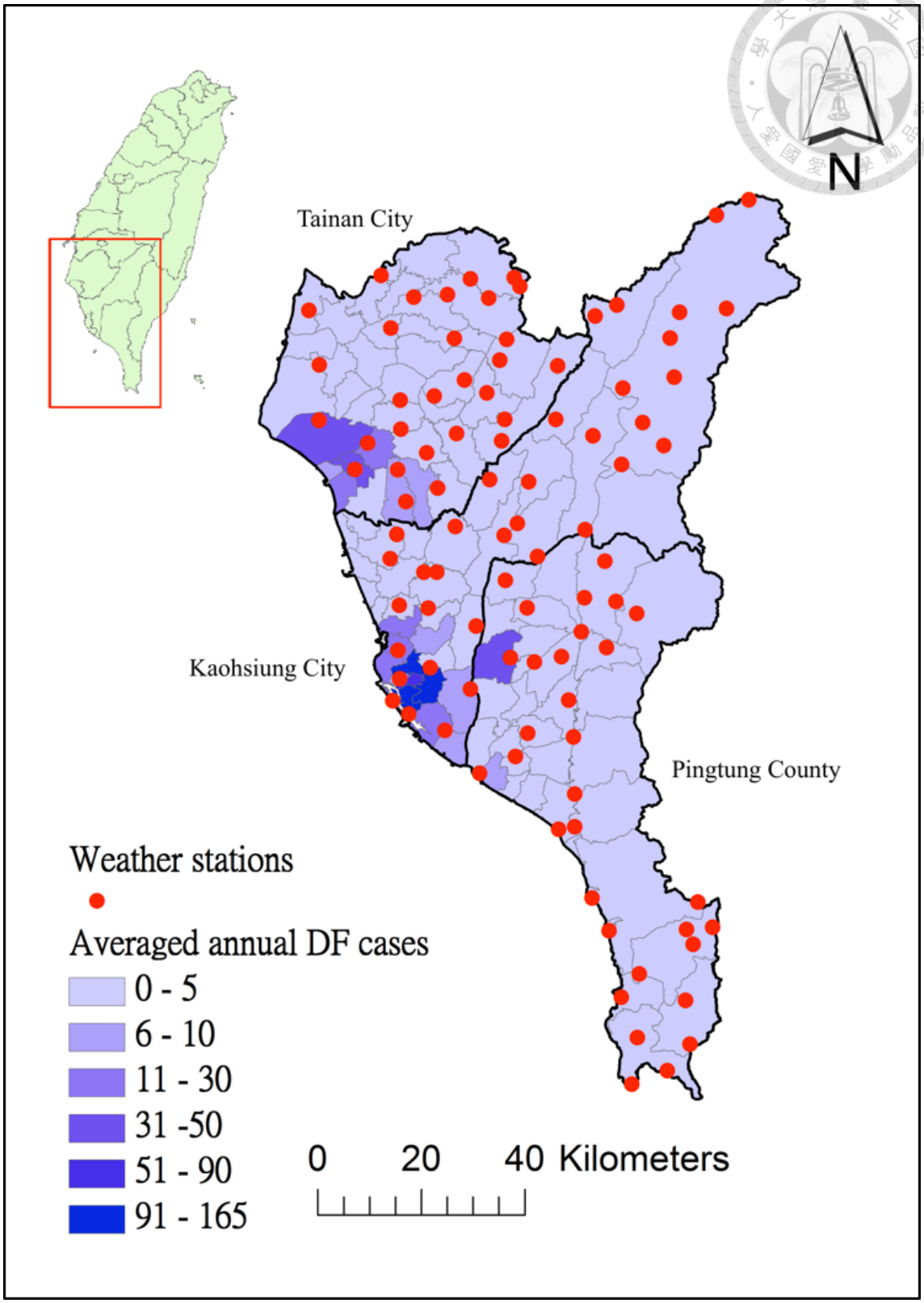


Figure 4.1: Map of the study area, which includes averaged annual DF cases of 107 districts in southern Taiwan, and the location of weather stations.

doi:10.6342/NTU202000011

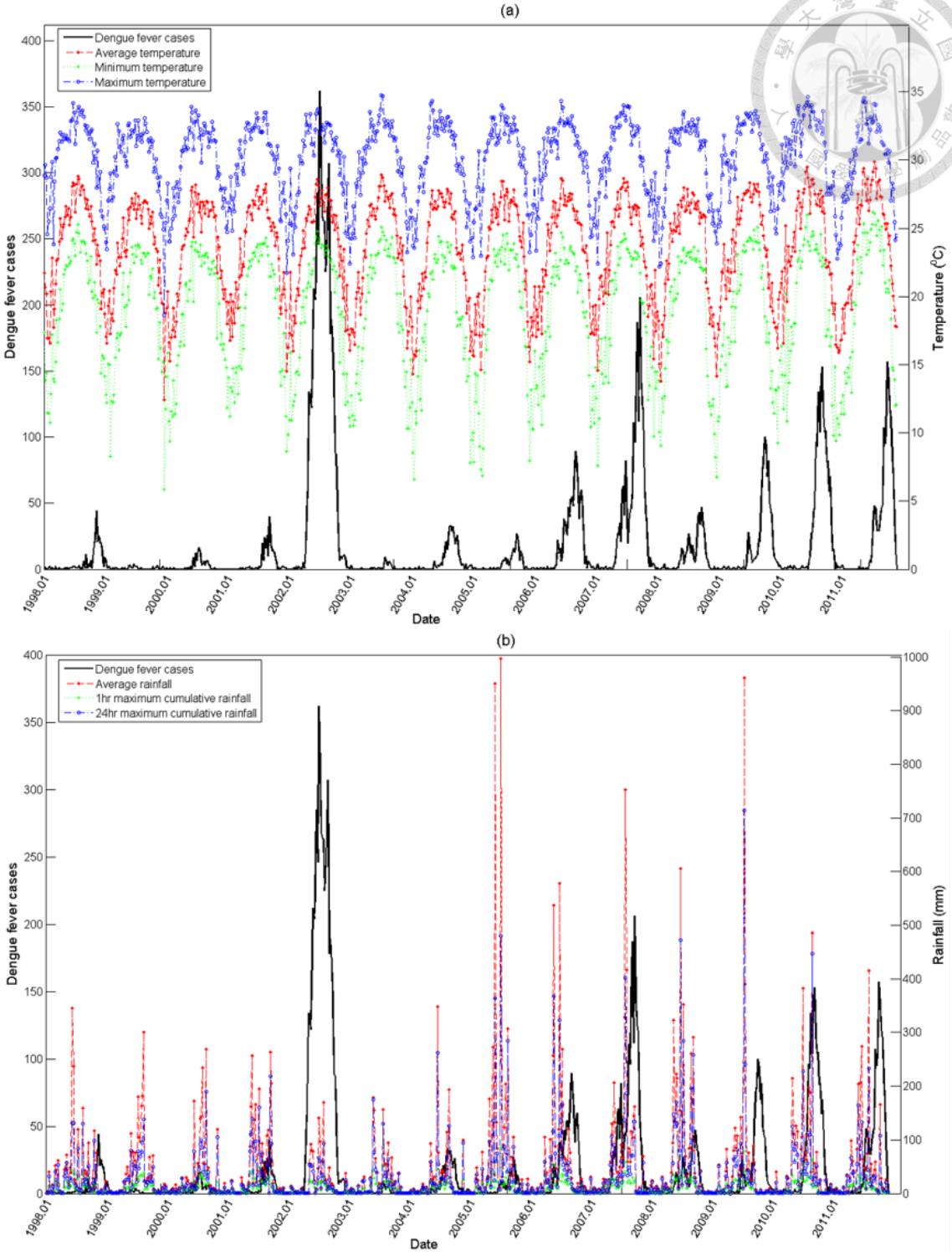


Figure 4.2: Trend plot of (a) weekly total dengue fever cases and temperature measures. Dengue fever cases(black); Average temperature(red); Minimum temperature(green); Maximum temperature(blue), and (b) weekly total dengue fever cases and rainfall measures. Dengue fever cases(black); Average rainfall(read); 1-hr maximum cumulative rainfall(green); 24-hr maximum cumulative rainfall(blue).

doi:10.6342/NTU202000011

(b) Site-specific or specificity knowledge (S- KB) that includes all knowledge bases (e.g., observations) that are specific to the region of interest and further split up into: hard data (i.e. exhibiting a satisfactory level of accuracy), and soft data that represent uncertainty in the measurements (including secondary information, imperfect observations, categorical data, and fuzzy inputs).

The BME method integrates the both knowledge bases, i.e. $K = G \cup S$, for the spatiotemporal estimation by the equations as follows[19, 21]

$$\begin{cases} \int d\chi(\boldsymbol{g} - \bar{\boldsymbol{g}} e^{\mu^{\mathrm{T}} \boldsymbol{g}}) = 0 \\ \int d\chi \xi_S e^{\mu^{\mathrm{T}} \boldsymbol{g}} - A f_K = 0 \end{cases} \tag{4.1}$$

where χ denotes space–time realizations of the dengue fever incidence distribution, $\boldsymbol{g}$ is a vector of g_α-functions ($\alpha = 1, 2, \ldots$) that represents stochastically the G-KB under consideration (the bar denotes statistical expectation), μ is a vector of μ_α-coefficients that depends on the space–time coordinates and is associated while also $\boldsymbol{g}$ (i.e., the μ_α express the relative significance of each g_α-function in the composite solution sought), the ξ_S represents the S-KB available, A is a normalization parameter, and f_K is the disease probability density function at each space–time point (the subscript K means that f_K is based on the blending of the core and site-specific KB). The $\boldsymbol{g}$ and ξ_S are the inputs in Eq. (4.1), whereas the unknown are the μ and f_K across space–time. In this study, the G-KB considers the space–time empirical relationships between the hydrological factors and DF incidences, which were analyzed by the DLNM method, and the space–time dependence among the disease incidences by using a spatiotemporal dependence function. The S-KB incorporates the DF observations with considering their associated uncertainties. The details will be described below.

4.4.2 Spatiotemporal DF modeling

The space–time spread of dengue fever epidemics can be influenced by a variety of climatic and non-climatic factors. This study characterizes the space–time variation of

doi:10.6342/NTU202000011

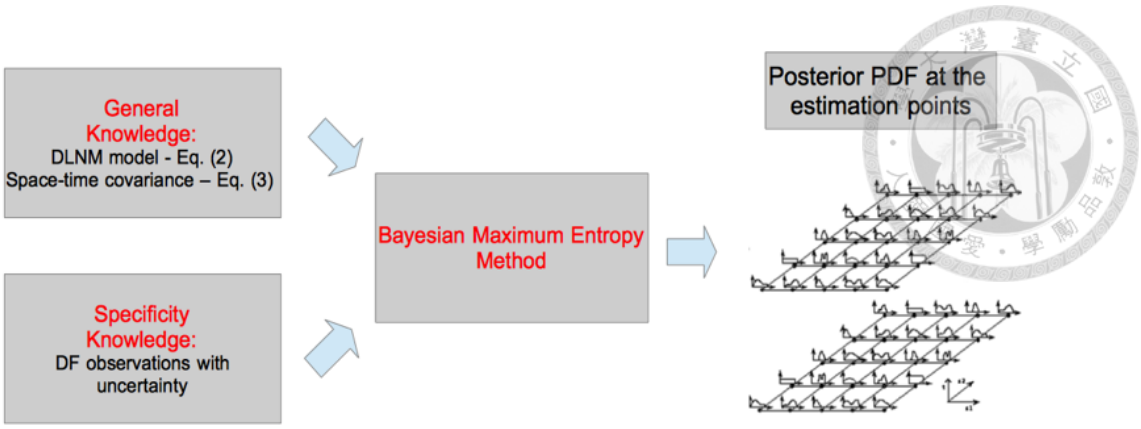


Figure 4.3: The conceptual flowchart of BME analysis in space-time DF modeling.

DF occurrences by integrating a generalized space–time trend model of DF incidence cases (considering seasonal and hydrological impacts), and structured spatiotemporal random effects (accounting for the differences between observed DF cases and estimated DF trend). This integration was based upon the BME framework. In other words, a S/TRF for space–time DF incidence is formulated by the G-KB of

1. DF trend model to provide the lead-time DF projections derived by the DLNM method to consider a variety of explaining variables, e.g., meteorological and hydrological factors

2. The space–time dependence model for the modeling of the non-climatic disease spread, e.g., human-vector interactions.

The flowchart of the BME framework of this study is shown in Figure 4.3 The details of spatiotemporal DLNM trend and dependence models consist of G-KB along with S-KB of dengue fever cases will be discussed in detail below.

This study investigated the climate-driven dengue fever variation by constructing an empirical relationships between the temporal variations of dengue fever incidences and hydrological factors, as part of G-KB. We applied the DLNM method to assess the nonlinear weekly lagged effects of selected weekly-based meteorological and hydrological

doi:10.6342/NTU202000011

variables, i.e., the averaged temperature, the maximum temperature, the minimum temperature, the total rainfall, the maximum 24-h rainfall, and the maximum 1-h rainfall. The details of DLNM analysis can refer to [38, 37]. The dengue fever incidences are assumed to be Poisson-distributed and can be denoted by $Zp = Z(s,t)\ Poisson(\mu_{s,t})$, where $\mu_{s,t}$ is the meteorological-based expected value of DF cases, and can be formulated by the DLNM model as follows

$$\log(\mu_{s,t}) = \alpha + \beta(\textbf{Year}) + f(\textbf{TP,lag}) + f(\log \textbf{RF}, \text{lag}) + f_T + f_S + \log(n_{s,t}) \quad (4.2)$$

where α is the intercept for the DLNM model and β is a vector of regression coefficients for the categorical variable; **Year** characterizing the cross-year variation of DF cases. f_T is a cubic spline function representing within-year relationship between disease incidences and time during the entire study period, i.e., seasonality which is assumed to be stationary from year to year. $f(\textbf{TP,lag})$ and $f(\log \textbf{RF}, \text{lag})$ are two cross-basis functions for temperature (**TP**) and rainfall (**RF**), respectively, that describe the space−time DF variations with respect to the interactions between the levels of temperature and rainfall variables, and their associated temporal lags. f_S represents the spatial random effects that characterize the general spatial variations across the study area[59]. $n_{s,t}$ is the population size at space−time location (s,t). The selection of significant hydrological factors for DLNM model is based upon the criteria of minimizing quasi-Akaike information criterion (QAIC) values. In the process of determining the most suitable temperature and rainfall cross-basis functions in the final model, only one temperature factor and one rainfall factor were considered in the variable selection process to prevent concurvity issues. For details of the variable selection process in DLNM model for DF analysis, refer to [18].

Though hydrological variables are closely associated with dengue fever occurrences, it should be noted that the disease transmission is highly complex which can also be influenced by other factors, e.g. the human and vector movements, virus serotypes, clustering infection, and dis- ease control interventions. For the purposes of characterizing the disease spread, a S/TRF was used for the modeling of the inconsistency between the observations and DLNM estimations of DF cases that represent the disease transmission

doi:10.6342/NTU202000011

not closely related to the influence of hydrological variations. The spatiotemporal dependence among the disease transmission is represented by the theoretical space–time nested covariance model c_X as below

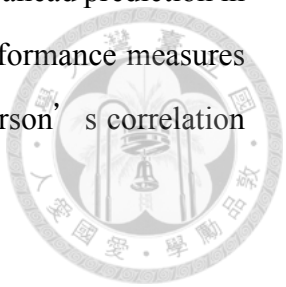
$$c_X(h,\tau) = c_1\boldsymbol{\delta}_{h_1,\tau_1} + c_2\exp(-\frac{3h^2}{a_{h_2}^2})(1-\frac{3}{2}\frac{\tau}{a_{\tau_2}}-\frac{1}{2}\frac{\tau^3}{a_{\tau_2}^3}) + c_3\exp(-3\frac{a_\tau h + a_h\tau}{a_{h_3}a_{\tau_3}}) \quad (4.3)$$

where $h = \mid s_i - s_j \mid$ and $\tau = t_i - t_j$ denote the spatial and temporal lags, respectively, between any pair of space–time points $p_i = (s_i, t_i$ and $p_j = (s_j, t_j)$. c_1, c_2 and c_3 are the sill coefficients, and $\boldsymbol{\delta}_{h,\tau}$ is the nugget component of the covariance model which accounts for the model uncertainties of the DLNM analysis. a_{h_i} and a_{τ_i} are the spatial and temporal autocorrelation ranges that characterize, respectively, the different covariance models nested in Eq. (4.3).

The G-KB accounts for the general space–time features of DF spread, i.e., how the hydrological variations can influence the DF incidences and how the DF cases are space–time auto-correlated across the entire study area. On the other hand, the S-KB consists of the differences between the observed and DLNM-observed DF data, which accounts for the spatiotemporal disease diffusion not explained by hydrological variations across different space–time locations. These differences can involve in various uncertainty sources, including the uncertainty of DLNM modeling and the uncertainty from the surveillance data of DF cases. The accuracy of DF observations can be highly influenced by the sampling scheme and frequencies. Some previous studies claimed that the underreporting is a prevalent issue in the surveillance of DF diseases[25]. To account for the high uncertainty of the DF data, the differences of DF data for BME modeling were assumed uniformly distributed with ranges from ± 3 cases at every space–time location. This study used the DF incidence data during the period of 1998–2011 to calibrate the G-KB, and, for the real-time prediction, the S-KB consists of uncertain DF observations in 2012. BME method stochastically assimilates the G-KB and S-KB by Eq. (4.1), and provides a "one-week-ahead" sequentially prediction over entire 2012 that is used to predict weekly DF cases at week n by given the weekly DF and hydrological data in the preceding weeks n - 1, n – 2, etc. To assess how DLNM and BME methods perform in the space–time distribution

doi:10.6342/NTU202000011

of DF spread, cross-validation analyses were performed for one-week ahead prediction in the entire 2012 for DLNM and BME methods respectively. The performance measures for cross-validation include root mean square error (RMSE) and Pearson's correlation coefficient.



4.5 Dengue fever diffusion modeling across space and time

The empirical relationship between the space–time variations of weekly-based DF incidences and a set of hydrological factors were assessed by the application of the DLNM model. Based upon the QAIC variable selection criteria, the weekly minimum temperature and weekly 24-h maximum cumulative rainfall were identified as the important hydrological factors, which are significantly associated with the space–time DF fluctuations. The identified associations reveal that the DF relative risks (RR) were related to the interactions between the levels of the two identified hydrological factors and their temporal lags. Among them, the RR of DF at a certain weather and temporal-delayed condition is defined as the ratio of the probability of DF occurrences under that specific condition to that under the reference condition[40]. In this study, the reference conditions for the weekly minimum temperature and weekly maximum 24-h rainfall are 18.82°C and 27.94 mm, which were the averaged levels of the two variables during the study periods, respectively. These associations for minimum temperature and logarithm of maximum 24-h rainfall are presented in terms of both a 3D-graph and a 2D contour plot, respectively, as shown in Figure 4.4. The temporal delayed effects of the weather variables were assessed up to 15 weeks, which was determined on the basis of the previous analyses[102, 17, 109].

Figure 4.4a shows a 3-D graph of the RR as it relates to minimum temperature and lags. This plot suggests that at greater minimum temperatures and more lagged weeks, a higher incidence of dengue fever was noted in southern Taiwan. The contour plot in Figure 4.4b is the contour representation of the 3-D graph that more clearly identifies the change in RR that occurs as the minimum temperatures and lags increase. Note that the RR gradually increased in all lagged weeks as the minimum temperature increases over 20°C, and reached the greatest RR = 2.86 (95 % CI = 2.62, 3.14) when the minimum tem-

doi:10.6342/NTU202000011

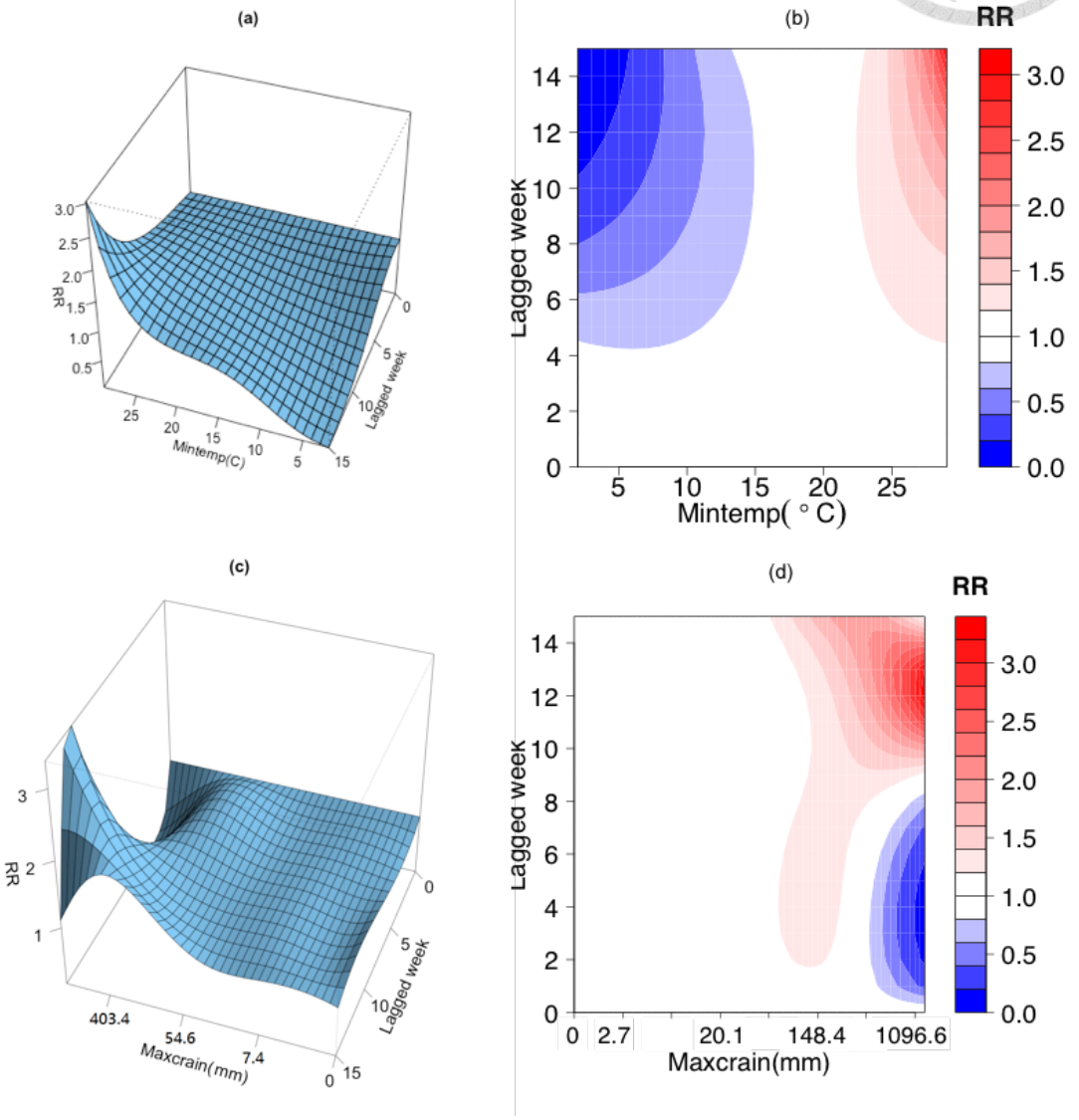


Figure 4.4: 3D graphs and their associated contour plots showing the relative risk of dengue fever incidence at lagged weeks corresponding to the weekly minimum temperature (a & b), and the logarithm of the weekly 24-hr maximum cumulative rainfall (c & d).

doi:10.6342/NTU202000011

perature climbed to 29°C at lagged week 15. Figure 4.4c displays how the RR varies with changes in the log of the maximum 24-h rainfall at different lags. Figure 4.4d clearly depicts that a RR was greater than 2 in every lagged week when the maximum 24-h rainfall was over 50 mm ($\approx e^4$). The RR gradually increased with greater maximum 24-h rainfall after lagged week 8, and eventually reached to the maximum value of 3.86 (95 % CI = 3.36, 4.44) when there was extreme rainfall at lagged week 12. Nonetheless, RRs were significantly lower than 1 for approximately 1 month when the extreme rainfalls, i.e., rainfall higher than 330 mm ($\approx e^{5.8}$), occurred. In summary, the DLNM model can explain 54.5 % space–time variation of dengue fever incidences, i.e., adjusted R-squared value. By the forward selection procedure, the explained variances by each covariate can be estimated. Among them, the nonlinear minimum temperature function explained the greatest proportion by 18.6% (f_S), the cross-year indicators explained 10.3% (β), the spatial function explained 10% (f(**TP,lag**), the yearly-invariant seasonal smoother explained 8.1% (f_T), and the nonlinear function of maximum 24-h rainfall explained the least proportion by 7.5% ($f(\log \mathbf{RF}, \text{lag})$).

Figure 4.5 shows the nested spatiotemporal covariance model characterizing the DF diffusion with distinct space–time scales, in which the three space–time separable models are comprised of Eq. (4.3) with sills c_1, c_2 and c_3 with values of = 2.41, 1.81, 0.6, respectively. The spatial correlation ranges for the Gaussian and exponential models are $a_{h_2} = 7$ and $a_{h_3} = 20$ in kilometers, respectively, and the temporal ranges for spherical and exponential models are $a_{\tau_2} = 13$ weeks and $a_{\tau_3} = 8$ weeks. Figure 4.6 shows the comparison between the observed DF cases and the predicted DF incidences by two spatiotemporal prediction approaches across the entire southern Taiwan in 2012. The two "one-week-ahead" DF prediction approaches are

1. DLNM method only accounting for the time-series observations of the hydrological variables.

2. BME method considering both the influence of hydrological factors by DLNM method and the real-time variations of spatiotemporal DF diffusion.

The comparisons show that the consideration of hydrological variation with DLNM method

doi:10.6342/NTU202000011

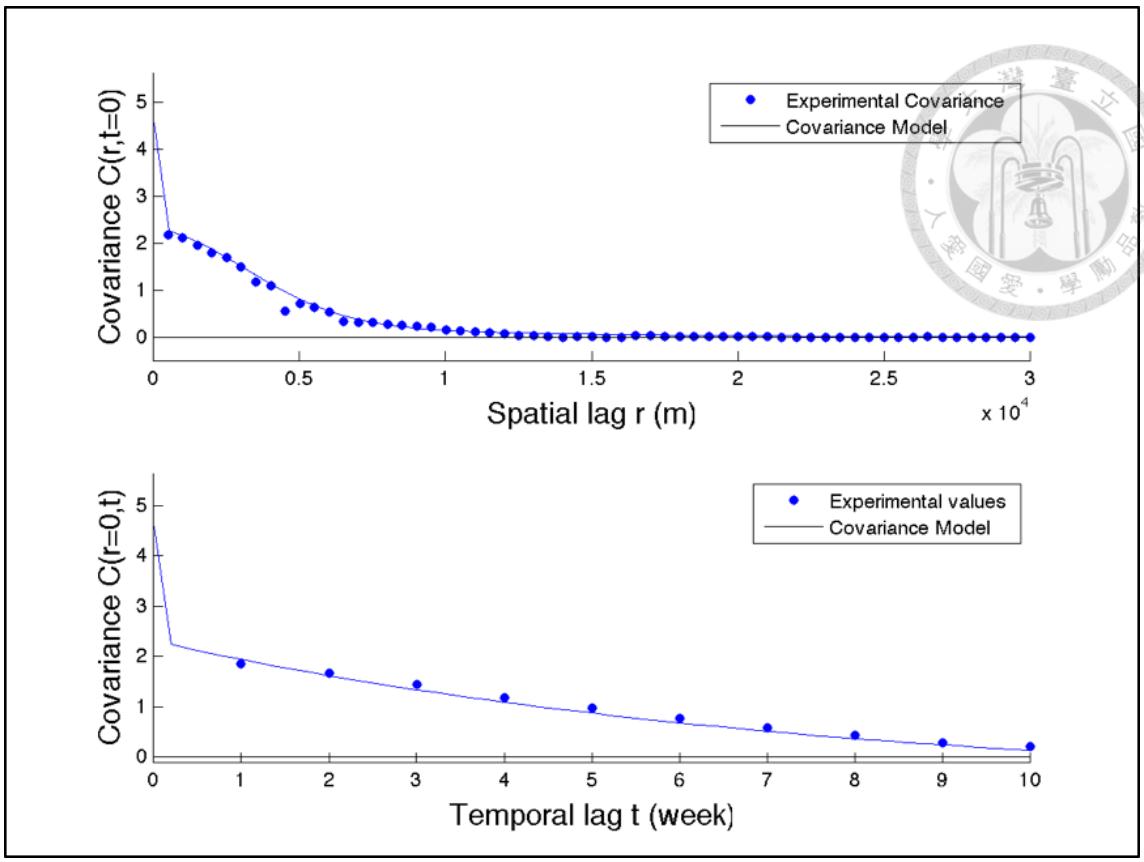


Figure 4.5: The nested spatio-temporal covariance model which characterizes the DF diffusion across space (top) and time (bottom).

can sufficiently provide the early warning messages about the DF epidemic, e.g. the initial and peak stage of the DF outbreak at around 15th and 37th week. The BME method improved the real-time prediction because it further considered the spatiotemporal autocorrelation among the disease observations resulting from the other factors, e.g. cluster infections and disease control interventions. To further reveal the performance of the spatiotemporal predictions of DF incidences, the time-series comparison of four townships with distinct DF variations were selected and shown in Figure 4.7, in which Figure 4.7a–d present the comparisons at Annan, North, Sanmin, and Lingya districts, respectively. Among them, the former two and the latter two districts are located in Kaohsiung and Tainan cities respectively. The spatial distributions of DF incidence rates at the selected weeks are shown in Figure 4.8, in which maps at the 39th and 40th weeks show the DF distribution during the peak of the 2012 outbreak.

The cross-validation results for both DLNM and BME methods are present in Table

doi:10.6342/NTU202000011

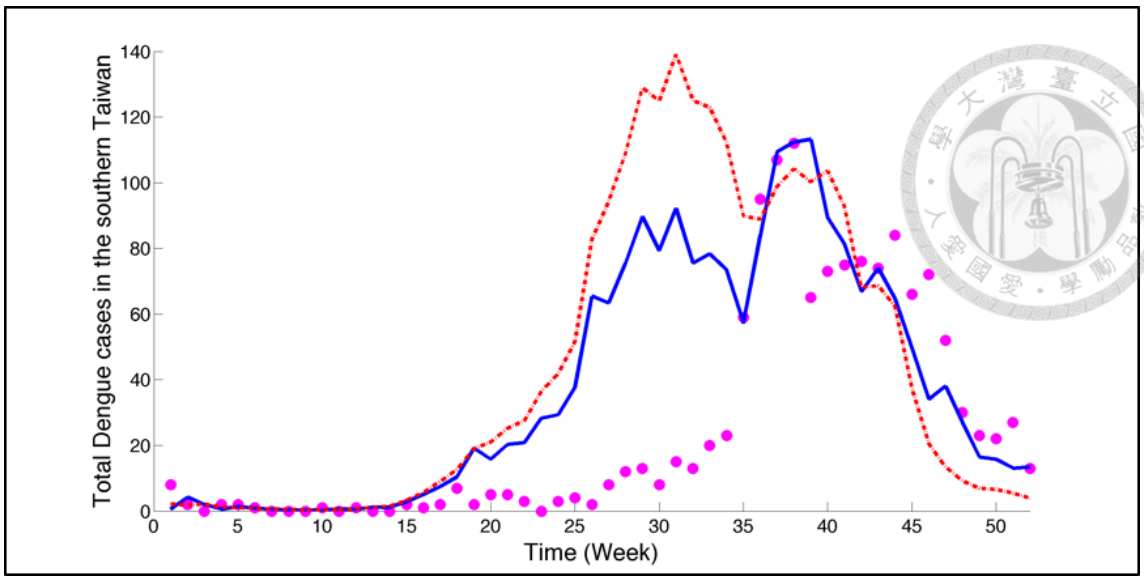


Figure 4.6: Comparison between DF cases: observed (dot) and predicted by the DLNM model (dashed line) and BME model (solid line) during 2012 in southern Taiwan.

Method	DLNM model (a)	BME model (b)	% Change from (a) to (b)
RMSE	12.89	6.89	-46.55
Pearson's correlation	0.19	0.51	68.42

Table 4.1: Comparison of the cross-validation results between weekly observed and predicted DF cases (DLNM: top; BME: bottom) at all townships in southern Taiwan during 2012

?? and Figure 4.9. Results show that the consideration of spatiotemporal DF diffusion in BME method can result in a significant reduction of RMSE, i.e., 46.55%, compared to DLNM-only model, and increase Pearson's correlation coefficient increased from 19% (DLNM) to 51% (BME). Both cross-validation measures indicate significant improvement of prediction accuracy by assimilating surveillance DF data with model prediction under BME framework.

4.6 Discussions

This study proposed a spatiotemporal DF early warning model based on a geostatistical framework, i.e. BME analysis. Due to its epistemic framework, BME method allows the consideration of the multi-sourced information and uncertainties for the modeling of the complexity of space–time disease, diffusion. BME has been widely used for the disease estimation under uncertainties; however, most of previous applications primarily investi-

doi:10.6342/NTU202000011

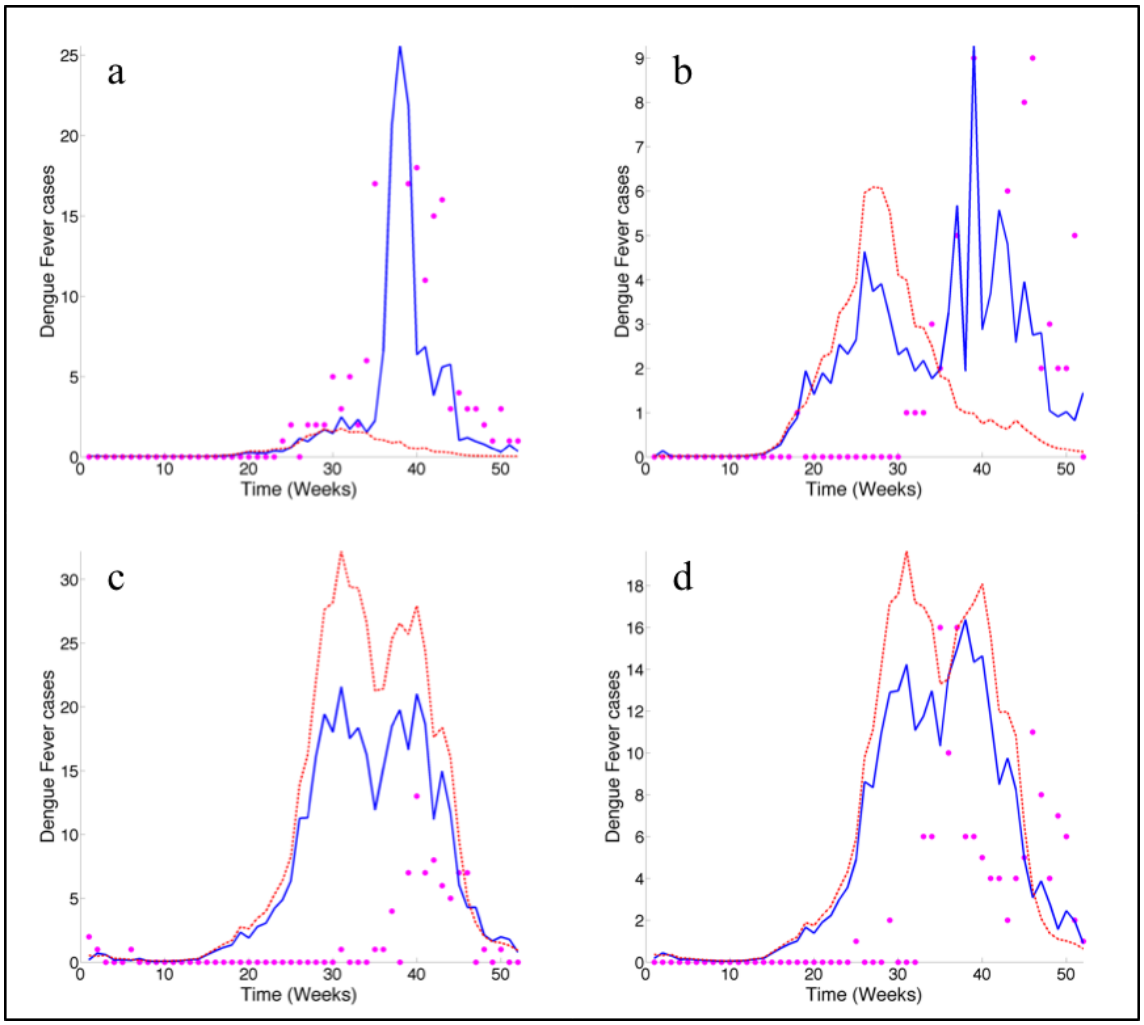


Figure 4.7: Comparisons between observed (dot) and predicted (DLNM: dashed line; BME: solid line) DF cases during 2012 at: a Annan District, b North District, c Sanmin District, and d Lingya District.

doi:10.6342/NTU202000011

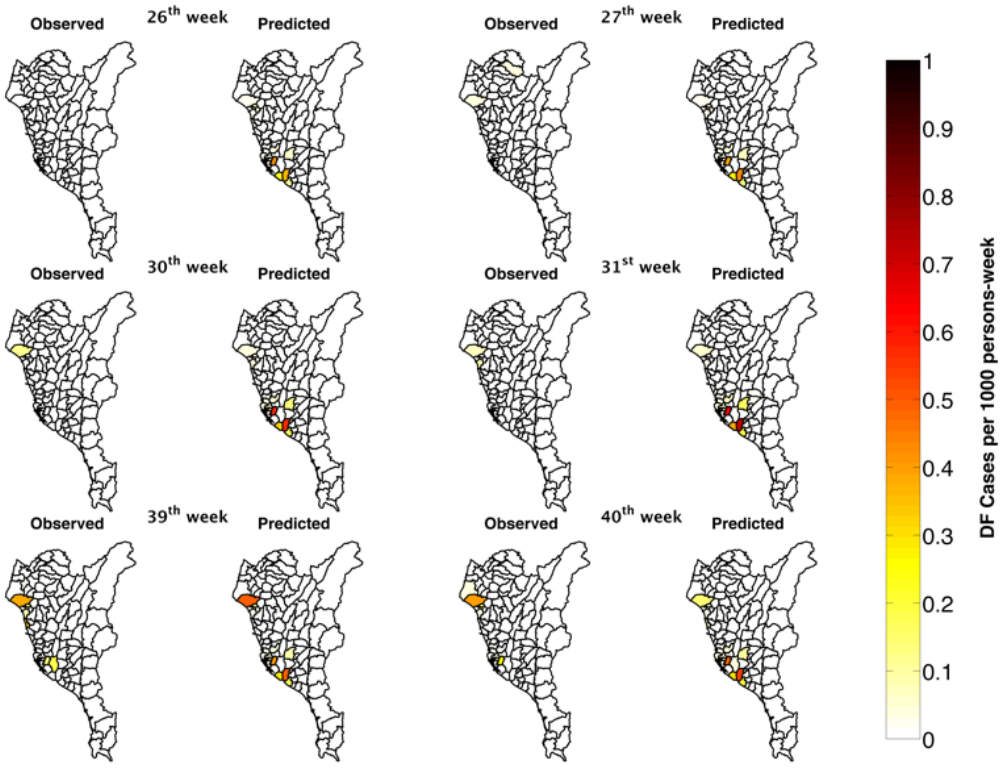


Figure 4.8: Spatio-temporal distributions of observed and predicted DF cases at the selected week of 2012.

doi:10.6342/NTU202000011

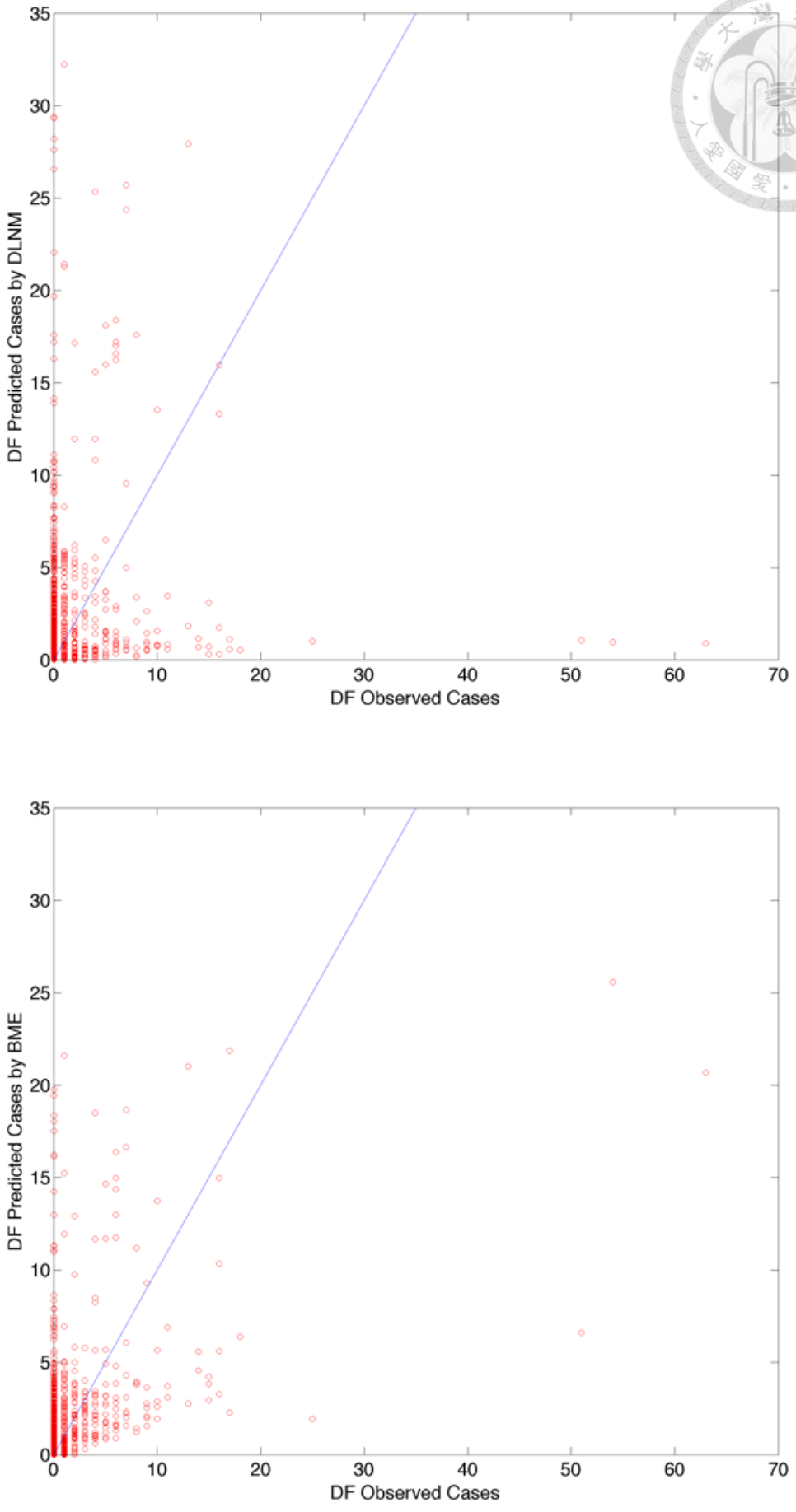


Figure 4.9: Comparison between weekly observed and predicted DF cases (DLNM: top; BME: bottom) at all townships in southern Taiwan during 2012.

doi:10.6342/NTU202000011

gated the spatiotemporal dependence among the disease observations without considering the influence of extrinsic factors to the disease spread[21, 107, 3, 2], which is important to vector-borne diseases, e.g. dengue fever. On the other hand, studies for the modeling of vector-borne diseases have considered climate-related variables as the major explanatory variables for temporal and spatiotemporal variations of the disease outbreak, i.e., considering climatic variables to estimate seasonal variation of the outbreaks; however, most of these approaches commonly do not account for the stochastic spatiotemporal associations of disease observations across the spatial and temporal domains[61, 67, 16, 69]. The proposed DF prediction model assimilated the knowledge bases accounting for the linear or nonlinear contributions from both temporal variation of external forcing, i.e., the impacts from hydrological variations, and spatiotemporal autocorrelation that can linearly or nonlinearly influence the spatiotemporal patterns of DF outbreak, i.e., space–time interactions, which result from the movements of human and vectors. In addition, our approach mitigated the multicollinearity among the time series of hydrological factors, which can result in the biased estimation of the statistical epidemiological associations.

Understanding the associations between vector-borne diseases and weather factors have been an important practice to establish the disease EWSs[61, 102, 67, 16, 58, 69, 68]. Determining the significant temporal lags between DF occurrences and their associated temporally varying contributors has been a major focus in previous studies[22, 67, 68, 109, 48]; however, it is difficult to assess the temporally-continuous lagged relationships due to the presence of seasonality in time series of both environmental and disease variables and the high collinearity within the time series of the hydro-climatic predictors. To account for the seasonality, a yearly-invariant nonparametric function was used in DLNM model to characterize the statistical relationships among the time series within the year. To address the high collinearity issue, the DLNM model included a varying lagged association across autocorrelated time series without linear assumptions, and identified that the weekly minimum temperature and maximum 24-h rainfall are significant to the DF risks in southern Taiwan. Results in Figure 4.4 show that a weekly minimum temperature greater than 20°C was associated with a dengue fever RR greater than 1, especially

doi:10.6342/NTU202000011

when the lag is over 4 weeks. Previous studies in southern Taiwan identified that temperature and DF incidences are highly associated at temporal lags from 1 month[92] to 2–3 months[102, 17, 109]. Our results show a monotonic RR increase as long as the minimum temperature is higher than the identified threshold temperature, i.e., 20°C. We also found that a higher minimum temperature decreased the lag period of a dengue fever outbreak, suggesting that higher temperatures may account for increasing rate of onset of the disease. This finding can be supported by the entomological evidences for DF vectors that the favorable temperature range for dengue fever transmission spans between 15 and to 35°C[102, 76]. A mathematical DF model in this study area showed the optimal average temperature for dengue fever transmission in the study area is about 28°C[16], which is similar to our results identifying the highest DF risk at the minimum temperature of the 29°C with 15 weeks lags. Furthermore, this study shows that the DF risk increases as soon as the weekly maximum 24-h rainfall exceeds approximately 50 mm. The increased risk for dengue fever can be seen from the onset of the rainfall and can last for at least 3 months. With extreme rainfalls, a rainfall of 330 mm of water or more, a dengue fever epidemic may be temporarily mitigated for 1 month. Our result is consistent with a previous empirical analysis for extreme rainfalls performed in the same study that the events of daily rainfall $\geq$ 130 mm can elevate dengue fever risk with the temporal lag about 70 days[15]. Our results further show the nonlinear associations between rainfalls and dengue fever, and that increased rainfall generally provides more favorable environments for mosquitoes and their larvae, and extremely heavy rainfalls can potentially have an adverse impact on the dengue fever vectors' habitats[65, 77, 86].

Hydrological factors can only partially explain the spatiotemporal dynamics of the DF spread. The DF transmission can also depend upon a variety of non-hydrological factors, including the number of infected and susceptible human hosts, the virus serotype, the transmission rate of virus among vectors and within human-vector interactions, as well as vector habitat conditions. For example, it has shown that the imported cases with favorable environmental condition can be an important factor to initiate a DF outbreak[83], and landuse patterns can highly associate with the spatial DF distribution[43, 97, 74].

doi:10.6342/NTU202000011

Though the hydrological-driven DF modeling by DLNM could provide a long-lead prediction for about 3 months in the study area, an early warning model should also be able to be adapted to the occurrence of DF observations. As shown in Figure 4.5, the spatiotemporal covariance characterizes the space–time DF diffusion with respect to human and vector movement. Among them, the space–time component with larger variance has spatial and temporal ranges of 7 km and 13 weeks. It implies that, for an epidemic year, the DF infections can generally stay in the same location for few months and extend to its neighborhood in the range of about 10 km. This short spatial correlation range can result from that strong disease control and intervention by the local government limit the transmission rate across the spatial locations. This diffusion characteristic provides the BME method to provide the ”one week ahead” prediction sequentially by accounting for data in real-time basis. Figure 4.7 shows the advantage of assimilating the real-time observations, i.e. providing the better estimations of the magnitude of DF incidences across space and time, in addition to the hydrological-driven warning of DF occurrence. The improvement is especially significant at locations where the dengue fever had never been observed in the past, e.g. Annan district shown in Figure 4.7a. In addition, the importance of data assimilation in the EWS development for vector-borne diseases has been only sparsely acknowledged in the existing literature[72, 109]. For the purpose of accounting uncertainty, the study assumed uniform distribution for DF data with ranges from ± 3 cases at every space–time location while the other study used ± 1 case[45]. However, using different ranges does not change the conclusion from the result of sensitivity analysis.

As Figure. 4.6 and 4.7 demonstrate, though the spatiotemporal distribution of DF predictions is consistent with the observed epidemic space–time patterns, the differences between the DF predictions and actual observations are still present. It should be noted that the diffusion and magnitude of DF epidemics commonly depend upon factors that introduce the nonlinear and drastic changes of the DF incidences, which can commonly not be characterized by stochastic models, e.g., DLNM. For example, disease cluster infections can significantly elevate DF cases in a very limited space–time extent, and the disease prevention implementation can impose different degrees of impact on transmis-

doi:10.6342/NTU202000011

sion rate across space and time. In addition, our proposed model predicts DF distribution on the basis of the spatiotemporal distributions of weather variables. In other words, the accuracy of our model can be defected by the interventions of other major factors, such as DF virus serotypes and disease control practices. Among them, virus serotypes have been considered to be an important factor for the cross-year variations of DF epidemic magnitudes, and disease control practices can diminish the magnitude of DF epidemics. The local government significantly increased their efforts in DF control, especially at the late spring. Particularly, in these 2 years, governmental agency conducted a new biological vector control practice in the study area, which may account for our over-prediction of DF epidemic in early stage of the outbreak, i.e., weeks 25–30, as shown in Figure 4.6 and 4.7. Despite of these limitations, this model can generate adequate warnings for either initial outbreak or the peak time of the DF epidemic with certain lead-time. In other words, the proposed model can be considered as a decision support approach for the disease prevention by providing the potential spatiotemporal patterns for the upcoming outbreak rather than providing accurate space–time DF incidence across the study area, because of the above mentioned unknowns and uncertainties. As shown in Figure 4.8, the predicted and actual spatiotemporal DF distributions in 2012 are similar. It implies that the DF prediction can detect the spatiotemporal hotspots of the epidemics, which are the important reference for the disease control agency, even though some deviation of predictions can present. This study only took the hydrological factors into consideration for the purposes of EWS. For the purposes of disease prevention, future study can further investigate the integration of additional DF-related factors, e.g. land use and imported case data, which can possibly improve the understanding the spatiotemporal distribution of DF risks.

doi:10.6342/NTU202000011

Chapter 5

An Efficient Spatiotemporal Data Calibration Approach for the Low-cost $PM_{2.5}$ Sensing Network: A Case Study in Taiwan

(*Published in Environmental International, 2019[63]*)

”生活袂曉過，萬事袂得過。天公伯仔你哪毋講話，敢係疼我疼甲傷超過，才會予我遮濟創治家己的機會。”

美秀集團 **Amazing Show**《生活袂曉過》

Low-cost sensors is enabled easy deployment and fast response in near real-time that are one of the most used supplements to achieve the two V's (*Volume* and *Velocity*) of *Big Data*. In environmental monitoring, low-cost sensors have been developed to measure for various fields including air pollution, flow rate, soil moisture and micro climate etc. Inevitably, sensors also come up with huge concerns about measurement uncertainty. This is a critical issue for environmental analytics, especially which is related to risk analysis.

doi:10.6342/NTU202000011

The most significant advantage of environmental sensors is to effective increase the resolution of environmental monitoring in space and time. The finer spatiotemporal resolution of environmental measurements is greatly helpful for better interpolation and calibration of environmental models. The models could get rid of neither scarcity nor sparsity problems due to limited monitoring stations which are relatively heavy and expensive. However, it is common sense that the measurements of low-cost sensors are more uncertain comparing with the standard measuring equipments. The neglect of the uncertainty of sensors will possibly cause deceptive results, in further, make change in public risk perception.

Taiwan government is promoting to build a intelligent environmental monitoring system in the present. In the vision of the government's plan, there will be thousands of low-cost sensors deployed at the entire island. Hence, how to ensure the quality and account the uncertainty of sensor measurements will be a challenge. Because of the large amount of data, the computational efficiency is considered as another very important factor. In this chapter, we will use the low-cost air pollution sensors as a example (PM2.5 in specific) that have been deployed in Taiwan to construct a calibration model. The model is expected to be generalized, efficient and easy to deploy with different computing environment.

5.1 Questionable IoT-based sensors as solution to air quality monitoring

Air pollution monitoring would also benefit from these new generation devices. By connecting a set of sensors, the monitoring network can gather and send air quality information to the community within a very short time[1, 4, 110]. The large data information can be used not only for mapping air pollution, identifying the sources, and tracking changes but, in further, predict extreme air quality events. Low-cost sensors provide an opportunity of constructing a ubiquitous and long-term monitoring system. With a finer resolution of air pollution map, the specificity and sensitivity of exposure assessment are expected to

doi:10.6342/NTU202000011

reduce significantly.

However, the large data volume is also accompanies by questions of data veracity and validity. How to ensure the quality of massive data has become a critical issue in data analysis. In general, the air quality monitoring data is used to quantify the adverse health effects and helps determine the regulatory policies. The data quality must meet the minimum requirements for reliability. According to U.S. Environmental Protection Agency, all air quality monitors are required to meet Federal Reference Method (FRM) and Federal Equivalent Method (FEM) criteria. Quality assurance (QA) and quality control (QC) of air quality monitoring data are strictly necessary to support legislative decision makings. These QA/QC protocols are established to maintain instrument performance that measure and report highly trustworthy values.

The increase in low-cost air quality sensor applications have resolved the spatial sparsity of the current regulatory air monitoring network and caught community organizations' attention. For the public health concerns, community groups have high demands on grabbing local information. Rapidly developing of low-cost sensors seems like a great chance for communities to understand and assess neighborhood exposures. For trust issues, low-cost air sensors are commonly compared to regulatory air monitors in communities. However, the unknown quality of low-cost air quality sensors give no assurance that measurements can be misleading due to inappropriate interpretation. U.S. EPA has indicated that precision and bias are the most concerns of low-cost air sensors [57]. The accuracy of sensors is very important to evaluate in terms of reliability. The performance of a low-cost air sensor evaluation become the first priority. Community Air Sensor Network (CAIRSENSE) compared the low-cost air quality sensors to the reference monitor which meet FRM/FEM criteria. The study shows that the selected $PM_{2.5}$ sensors exhibited poor correlation with the reference monitor ($0.33 \leq r^2 \leq 0.4$) and other gaseous pollutants had stronger correlation ($0.57 \leq r^2 \leq 0.94$) [55]. Mukherjee et al.[75] examined the performance of particulate matter sensors under real-world conditions and showed the sensors could report relative low measurements to the well-characterized instruments. Other evaluations of PM sensors also show widely varying range of correlation with FRM/FEM

doi:10.6342/NTU202000011

measurements ($0 \leq r^2 \leq 0.8$) [96].

Sensor performance evaluations have pointed out the inconsistency between low-cost sensors and high-quality regulatory instruments. The variances of low-cost sensors measurement influence on health risk perception. It is important for public to recognize the difference between low-cost air quality sensors and the regulatory monitoring stations. To minimize sensor uncertainties, calibration is essential for producing validated data. U.S. EPA suggests in the guidebook that calibration is needed if at all possible and calibration should be processed before, during and after the data collection [100]. The response of low-cost sensors would be compared to the response of a reference instrument. The air quality monitoring stations that meet QA/QC standards, e.g., FRM and FEM sites, are commonly used as the reference instruments. The purpose of comparison procedure is to construct a relationship curve of low-cost sensors and reference standards. In field calibration, researchers found the sensor performance would vary because of the changes of meteorological conditions. The studies included mainly meteorological factors such as humidity, temperature, and cross-sensitivity into the calibration process accounting for influences on gaseous sensors [10, 30, 24, 70, 71, 49]. Likewise, meteorology is also considered as confounding factor to affect the response of particle sensors. Instead of measuring particulate mass directly, PM sensors use light-scattering method to count particles that pass through the optical cell. In CAIRSENSE study, incorporation of artifacts such as temperature and humidity contributed minor increases of agreement between some particles sensors and FEM.

Taiwan Air Quality Monitoring Network (TAQMN) is the largest gas pollutants monitoring network in Taiwan which is operated by Taiwan Environmental Protection Administration (TWEPA). For the purpose of air pollution control, TWEPA have built TAQMN and monitored air quality over 20 years. Although TWEPA have monitored air pollutants through TAQMN, the spatial and temporal resolution are still very sparse. For public health concerns, the information provided by TAQMN is still not enough. In recent years, specific low-cost air quality sensors which target $PM_{2.5}$ have been developed by the local maker community of Location Aware Sensing System (LASS) in collaboration with

doi:10.6342/NTU202000011

scientists, governmental partners and industrial companies. The low-cost $PM_{2.5}$ sensors measurements can be access via the designed open data platform [14]. Therefore, the measurements of low-cost sensors are frequently compared to the measurements of TAQMN in communities and result in the issue of information inequality. To avoid the public receiving and misunderstanding uncertain measurements of low-cost air quality sensors, it is urgent to develop a calibration procedure in Taiwan.

5.2 Applications of Commercial $PM_{2.5}$ sensors and regulatory air quality stations in Taiwan

Currently, there are seventy-six air quality monitoring stations operated by TWEPA throughout the island. These stations constantly collect six criteria pollutants including sulfur dioxide (SO_2), nitrogen oxides (NO_x), ozone (O_3), carbon monoxide (CO), PM_{10} (particles less than 10 micrometer in diameter) and $PM_{2.5}$. Among them, there are five stations located within Taichung metropolitan area, which is the second largest city in Taiwan with the population size of 2.79 million in 2018. Recently, awareness of $PM_{2.5}$ health issues in Taichung is rapidly raising because the largest coal-fired power plant in the world is located in the city. This study used the hourly $PM_{2.5}$ from the TWEPA stations in Taichung.

Since the early 2016, the popularity of low-cost air quality sensors has been rapidly increased island-wide, particularly in metropolitan areas. Over 5,000 air quality sensors have been installed in the last two years in Taiwan. In Taichung, local communities have volunteered to install low-cost $PM_{2.5}$ sensors of various types over the most of highly urbanized parts of the study area. Among these low-cost sensors, the majority belongs to the AirBox air quality detection system produced by Edimax Inc[28], that currently account for more than 85% of the all recorded low-cost sensors. The popularity of this particular sensor primarily resulted from its high accessibility and its manufacturer promoting strategy. To have the same variability and stability of sensing quality, this study only considers $PM_{2.5}$ observations from AirBox in our analysis. There are total 438 AirBox devices have been used in this study. The spatial distribution of AirBox sensors and TWEPA stations

doi:10.6342/NTU202000011

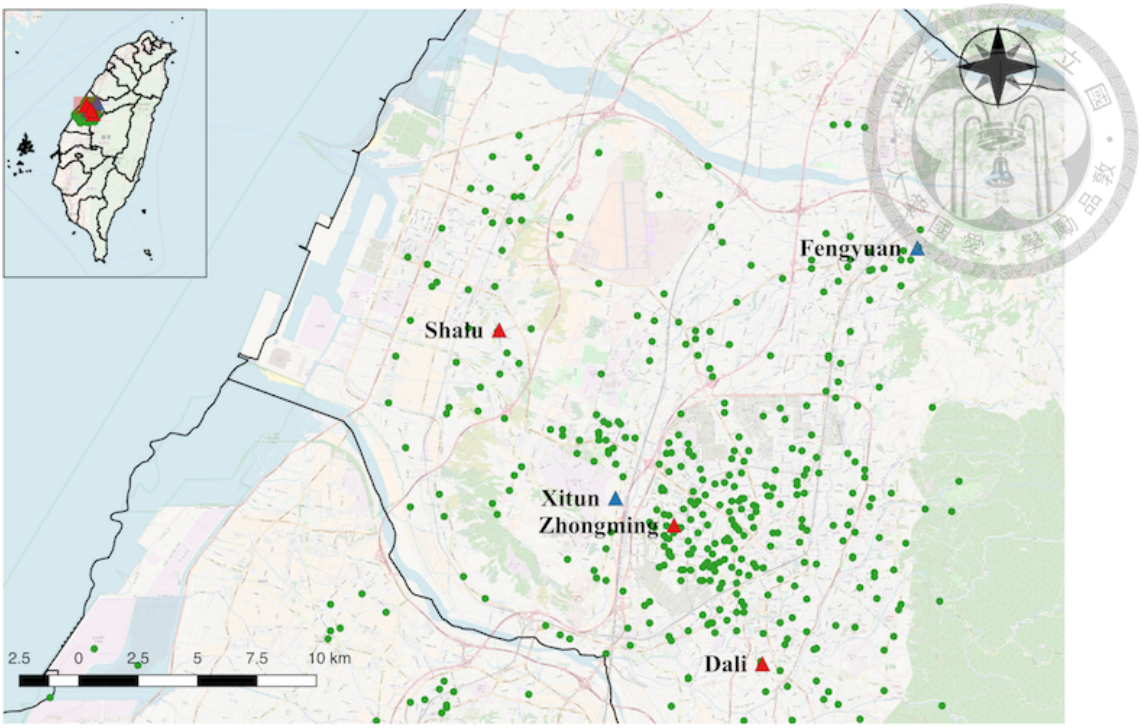


Figure 5.1: Spatial distribution of TWEPA stations (triangles) and AirBox devices (solid circles) in Taichung metropolitan area, where the red triangles are the places with both sensors collocated.

in Taichung is shown in Figure 5.1.

In general, the AirBox reports $PM_{2.5}(\mu g/m^3)$ observations roughly every 5 minutes, along with some environmental variables, i.e., temperature(°C), and relative humidity(%); namely, the observation times and intervals are not aligned among the low-cost sensors. In order to understand the observation quality of low-cost sensors, AirBox were deployed at Zhongming, Shalu and Dali stations respectively to perform the concurrent observations along with the regular monitoring instruments at TWEPA stations, as shown in Figure 5.1. In addition, to assess the consistency of the AirBox observations, two AirBox sensors are installed at the three concurrent observation locations with TWEPA stations. The observations used in our analysis are in the period between November 1st and December 31st, in 2017. Because of the different temporal resolutions between the observations from the two monitoring systems, for the purposes of comparison, the $PM_{2.5}$ measurements of AirBox were aggregated into hourly-averaged data.

Table 5.1 shows the summary statistics of the $PM_{2.5}$ of both AirBox and TWEPA sta-

doi:10.6342/NTU202000011

$PM_{2.5}$ observations ($\mu g/m^3$)										
Stations	TWEPA		AirBox-1				AirBox-2			
	Mean	Std	Mean	Std	25%	75%	Mean	Std	25%	75%
Dali station	18.75	13.59	24.67	12.5	16.5	31.12	23.67	13.57	15.84	28.17
Shalu station	18.29	15.82	16.95	9.66	10.2	22.45	19.11	10.77	12.2	24.0
Zhongming station	22.01	17.1	13.79	10.54	5.5	21.0	-	-	-	-

Table 5.1: The summary statistics of the $PM_{2.5}$ observations from the TWEPA and AirBox sensors.

tions at the three selected locations during the study period. It shows that the averaged $PM_{2.5}$ were different among the stations. The time series comparison and the distribution of differences between $PM_{2.5}$ measurements of AirBox sensors and TWEPA stations are shown in Figure 5.2 and Figure 5.3, respectively. From the raw measurements as shown in Figure 5.2, the temporal variations in all AirBox are similar to the reference stations. Consequently, low-cost sensor is able to capture the trends of $PM_{2.5}$. However, the AirBox devices mainly observed higher $PM_{2.5}$ concentrations than the TWEPA stations. These figures indicate that the observations among the AirBox themselves are relatively similar during the study period; however, compared to those from the TWEPA stations, the $PM_{2.5}$ observations at AirBox devices were averagely greater at least 13 $\mu g/m^3$, even over 20 $\mu g/m^3$ at Dali station. In other words, the $PM_{2.5}$ observations at AirBox sensors have systematic biases compared to the TWEPA stations, i.e., regulatory monitoring stations, and these biases can vary across space and time.

5.3 Space-time anomaly detection processes

The data calibration approach for the low-cost $PM_{2.5}$ observations can be divided into three parts, i.e., data cleaning, bias modeling, and uncertainty assessment. In the data cleaning process, the observations that present the inconsistency with the others in space and time are considered to be outliers and will be removed in advance to the modeling of the bias between low-cost and regulatory observations, because the presence of the outlier can heavily affect the parameter estimation of the bias relationship modeling. To identify the space-time outliers, this study considered the outliers in three different contexts, i.e., spatial outlier, temporal outlier, and space-time outlier. Spatial/temporal outliers are the

doi:10.6342/NTU202000011

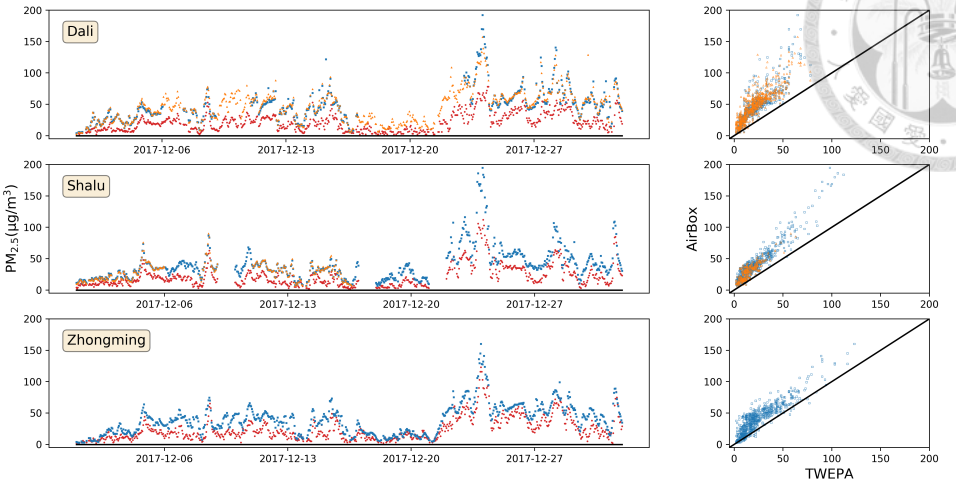


Figure 5.2: The raw data comparison of AirBox and TWEPA stations at three selected sites in December, 2017. Left column is time series comparison and right column is scatter plots of $PM_{2.5}$ observations from AirBox vs. TWEPA stations. Red dot: TWEPA observations; Blue square: AirBox-1; Orange triangle: AirBox-2.

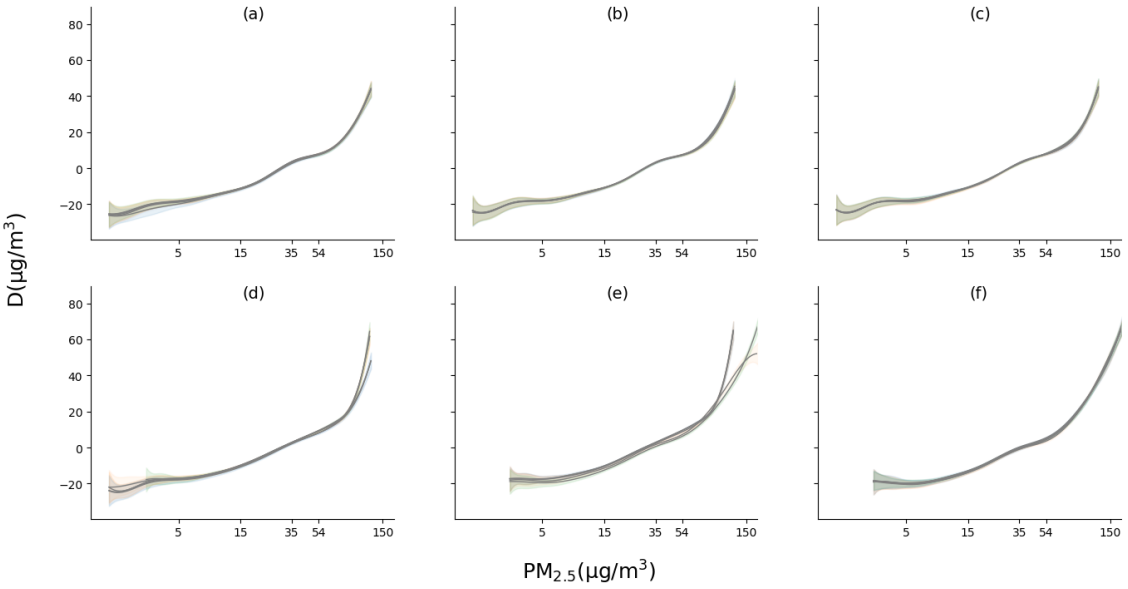


Figure 5.3: The daily bias correction relationships in Eq. (5.3) with respect to AirBox $PM_{2.5}$ observations on (a)12/1-12/5, (b)12/6-12/10, (c)12/11-12/15, (d)12/16-12/20, (e)12/21-12/25, and (f)12/26-21/31, respectively.

doi:10.6342/NTU202000011

observations which is significantly different from their spatial/temporal neighborhood. The spatial/temporal neighborhood is a pre-specified window or bandwidth in which the observation is situated in the center of this space/time area. In data cleaning process, the data low-cost sensor are only considered.

An observation, $Y(\mathbf{p}_i)$, can be considered to be an outlier while its value is significantly departing from its expected value in space, time, or space/time neighborhood. The expected value is estimated by the inverse distance weighting (IDW) interpolation method, and can be expressed by a weighted average of the observations in its neighborhood, as below

$$\begin{cases} \mathbf{E}[Y(\mathbf{p}_i)] = \frac{\sum_{j=1}^{N} \omega(\mathbf{p}_j) Y(\mathbf{p}_j)}{\sum_{j=1}^{N} \omega(\mathbf{p}_j)}, \text{ if } \|\mathbf{p}_i - \mathbf{p}_j\| < \gamma \\ \omega(\mathbf{p}_j) = \frac{1}{\|\mathbf{p}_i - \mathbf{p}_j\|^m} \end{cases} \tag{5.1}$$

where $\mathbf{p}$ is the location vector and can be seen as $\mathbf{p} = (\mathbf{s}, t)$. $\omega()$ is the weighting function that is commonly decreasing as the distance between the estimated location and its neighborhood increases. m is a pre-specified value that defines its decreasing behavior and this study considered it to be 2. γ is the size of bandwidth used for spatial or temporal windows. N is the number of observations within the space/time windows of concern. To determine the outliers, the following criteria is used

$$\alpha \cdot \sigma_\gamma + \mathbf{E}[Y(\mathbf{p}_i)] < Y(\mathbf{p}_i)$$

or

$$Y(\mathbf{p}_i) < \mathbf{E}[Y(\mathbf{p}_i)] - \alpha \cdot \sigma_\gamma$$

where σ_γ is the standard deviation of $PM_{2.5}$ within the neighborhood. α is a parameter to determine the threshold where is fixed at 3 in this study. The proposed outlier detection procedure is used for pure spatial and pure temporal cases, respectively, to identify abnormal observation in both spatial and temporal contexts. In other words, every observation can be identified to be a spatial outlier, a temporal outlier, or both. Because the abnormal values can be either outliers or extreme events, e.g., a short-period pollution event,

doi:10.6342/NTU202000011

in order to avoid the removal of true extreme values, the space-time outliers are only removed in this study, i.e., the outliers that are considered to be spatial and temporal outliers concurrently.

5.4 Nonlinear modeling for the biases from low-cost sensors

For the purposes of data calibration for the observations from low-cost sensors, we defined the observation bias to be $D(\mathbf{p}) = Y(\mathbf{p}) - Y_{\text{Ref}}(\mathbf{p})$, where $Y(\mathbf{p})$ and $Y_{\text{Ref}}(\mathbf{p})$ are observations from the low-cost sensors and TWEPA monitoring stations at location $\mathbf{p} = (\mathbf{s}, t)$. The Generalized additive model (GAM) model is used to assess the relationship between the observed biases and the environmental factors. GAM is a nonparametric regression method that can obtain the nonlinear relationships of dependent variables and their covariates of concern. The environmental factors used in our study include both meteorological conditions and air quality levels. As for the meteorological variables, temperature, and humidity are considered because they are the only meteorological variable readily available in every low-cost sensors, though these observations can be uncertain. The observations of all criteria pollutants at TWEPA stations were all considered in this analysis. In addition, the $PM_{2.5}$ levels observed by the low-cost sensors are also considered to be a covariate for the bias modeling. To obtain the optimal model, this study performed a forward selection process with Bayesian Information Criterion(BIC) values for the model selection. The observed biases can therefore be characterized by a GAM relationship as follows

$$g(\mu) = \beta + f(\mathrm{T}) + f(\log \mathrm{PM}_{2.5}^{\text{AirBox}}) + f(\log \mathrm{PM}_{2.5}^{\text{TWEPA}}) \tag{5.2}$$

where g is the Gaussian link function and $\mu = \mathbf{E}(D(\mathbf{p}))$ is the expected value of the observed bias at space-time location, $\mathbf{p}$. β is the intercept for GAM. $f(\mathrm{T})$ is the P-spline smoothing functions characterizing the effects of the temperature on the biases. The $PM_{2.5}$ levels obtained from both AirBox and TWEPA are shown their significant associations to the observed biases. Among them, the relationship with $PM_{2.5}$ observation itself is charac-

doi:10.6342/NTU202000011

terized by $f(\log \mathrm{PM}_{2.5}^{\mathrm{AirBox}})$. The use of logarithm scale for $PM_{2.5}$ is to reduce the leverage effect from few extreme $PM_{2.5}$ observations. Similarly, the $PM_{2.5}$ level at TWPEPA station is also considered in $f(\log \mathrm{PM}_{2.5}^{\mathrm{TWEPA}})$, while it is primarily used for overfitting mitigation while extremely high $PM_{2.5}$ is present. It should be noted that the $PM_{2.5}$ levels from the two types of sensors should have high correlation to each other. To avoid the collinearity effect, the formulation of Eq. (5.2) can essentially be decomposed into two layers shown below

Layer 1:

$$g(\mu_{var}) = \beta_{var} + f(\log \mathrm{PM}_{2.5}^{\mathrm{AirBox}}) + f(\mathrm{T}^{\mathrm{AirBox}}) \tag{5.3}$$

Layer 2:

$$g(\mu_{res}) = \beta_{res} + f(\log \mathrm{PM}_{2.5}^{\mathrm{TWEPA}}) \tag{5.4}$$

where $\mu_{var} = \mathbf{E}(D)$ is the expected variations of the observation biases. This model primarily accounts for the observation biases associated from the environmental factors. $\mu_{res} = \mathbf{E}(D - \mathbf{E}(D))$ is the expected value of fitted residuals from Layer 1 model. β_{var} and β_{res} are the two intercepts for the two GAM layers above, respectively. The final data calibration model requires the $PM_{2.5}$ from the both AirBox and TWEPA sensors. For those sensors not located at the same place as TWEPA stations, this study used IDW method to quicky estimate the pseudo-TWEPA measurement in hourly basis. This estimation can possibly induce the uncertainty to the calibration process.

To assess the robustness of the proposed bias adjustment model, this study designed a cross-validation procedure. There are two types of uncertainty were examined. First, the spatial consistency of bias relationship is investigated from applying a calibration model obtained from a specific location on the calibration at the other stations with concurrent observations. Second, the uncertainty of calibration models introduced from space-time estimations of the TWEPA $PM_{2.5}$ levels is checked by comparing the two calibration results that used the observed and estimated TWPEA $PM_{2.5}$ levels respectively. The Root Mean Square Error (RMSE) is used for the calibration performance assessment.

doi:10.6342/NTU202000011

5.5 The biases relationship between reference stations and $PM_{2.5}$ sensors

This study performed a dynamic calibration process in daily basis that the bias modeling for a specified date used the data collected within 30 days prior to that specific date. For example, the calibration model of December 1 used the data from November 1 to November 30 for the analysis. Hence, with the dataset from November 1 to December 31, 2017, we constructed 31 daily data calibration models of the form of Eq. (5.2) over the entire month of December. Figures 5.4-5.6 show the daily results of nonlinear relationship of observation biases, i.e., μ with respect to the three significant covariates, i.e., AirBox $PM_{2.5}$ and T measured at AirBox, and the $PM_{2.5}$ observations at TWEPA stations. All of these 31 daily relationships were based upon the data the three TWEPA stations with multiple sensors operating simultaneously. For the purposes of better presentation, the 31 nonlinear relationships with each covariate are grouped every five consecutive days and plotted separately. The results shows that the daily data calibration relationships with respect to the three covariates are mostly consistent during the study period. In general, dissimilarity among the daily calibration results may occur as the data ranges were changed significantly in a short period, such as the presence of extreme $PM_{2.5}$ observations. Figure 5.4 shows the observation bias is clearly $PM_{2.5}$ dependent. The bias increases exponentially as $PM_{2.5}$ level in AirBox increases. Figure 5.5 identified the negative associations with the observation biases and the temperature in AirBox. More specifically, the significant negative bias correction for $PM_{2.5}$ observations in AirBox is required while its temperature is higher than 26 °C, and this correction is amplified as the temperature increased.

Figure 5.6 presents the bias correction relationship should be negatively associated with the $PM_{2.5}$ in TWEPA stations derived from Eq. (5.4). Results show that Eq. (5.4) can become significantly effective while the $PM_{2.5}$ at TWEPA station are greater than about 35 $\mu g/m^3$. It should be noted that the scales of $PM_{2.5}$ in Figure 5.4 and Figure 5.5 are based upon AirBox and TWEPA observations, respectively. The nonlinearly increasing bias association can affect not only the bias but also the scale, i.e., $PM_{2.5}$ covariate, in

doi:10.6342/NTU202000011

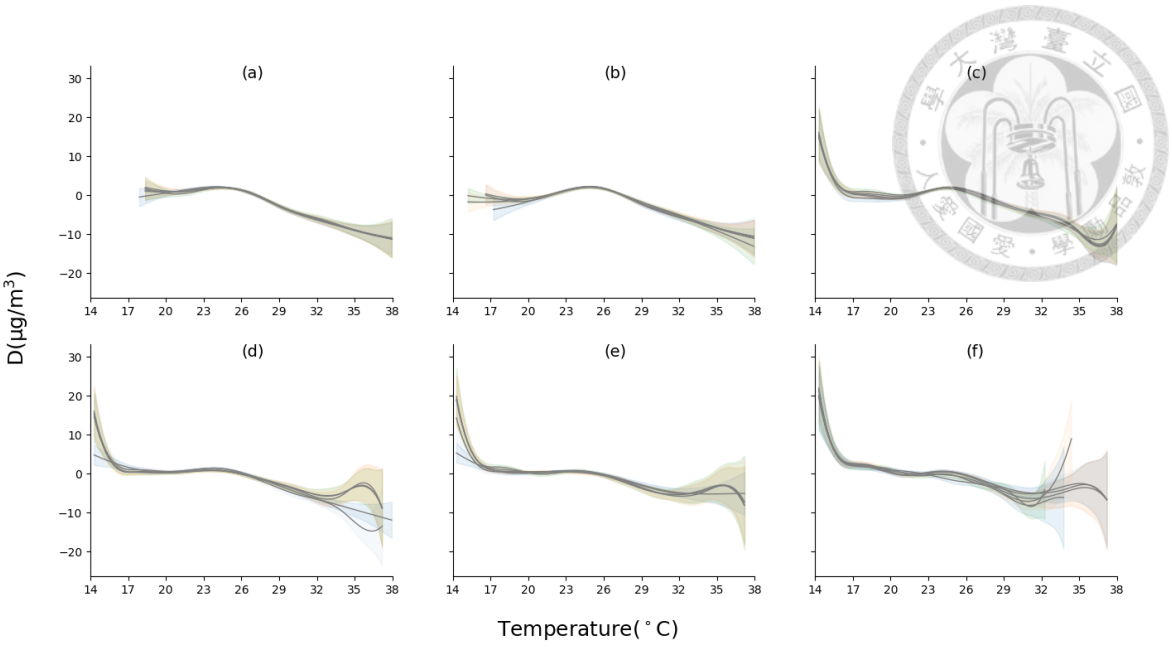


Figure 5.4: The daily bias correction relationships in Eq. (5.3) with respect to AirBox temperature observations on (a)12/1-12/5, (b)12/6-12/10, (c)12/11-12/15, (d)12/16-12/20, (e)12/21-12/25, and (f)12/26-21/31, respectively.

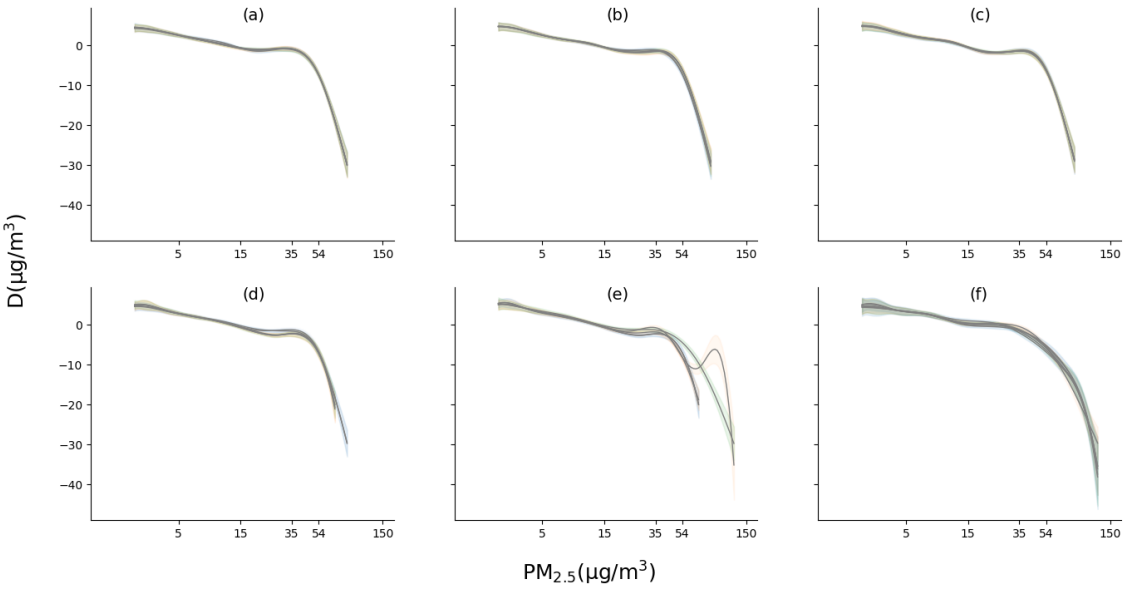


Figure 5.5: The daily bias correction relationships in Eq. (5.4) with respect to TWEPA observations on (a)12/1-12/5, (b)12/6-12/10, (c)12/11-12/15, (d)12/16-12/20, (e)12/21-12/25, and (f)12/26-21/31, respectively.

doi:10.6342/NTU202000011

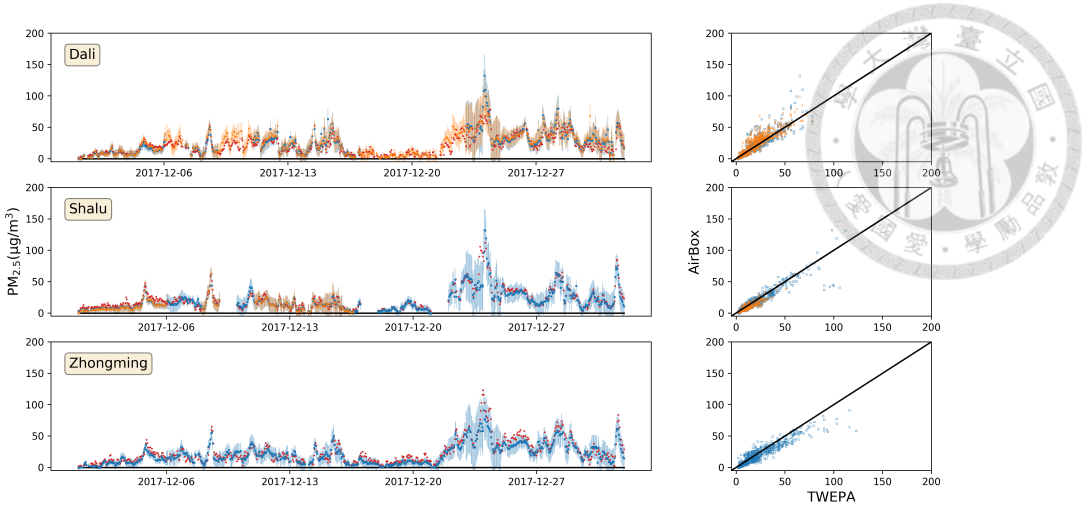


Figure 5.6: The calibrated results of AirBox and TWEPA stations at three selected sites in December 2017. Left column is time series comparison and right column is scatter plots of $PM_{2.5}$ observations from AirBox vs. TWEPA stations. Red dot: TWEPA observations; Blue square: AirBox-1; Orange triangle: AirBox-2. (For inter- pretation of the references to colour in this gure legend, the reader is referred to the web version of this article.

Layer 1 model and therefore the overfitting behavior of Layer 1 model is observed with high $PM_{2.5}$ concentrations.

After the calibration process applied, the results shown in Figure 5.7, the calibrated measurements of AirBox are close to the reference standards. The results show RMSE decreases after calibration from 31.34 and 29.73 to 9.66 and 8.29 at Dali station for AirBox-1 and AirBox-2, respectively. RMSE of Shalu station decreases from 15.55 and 22.48 to 4.88 and 7.34 for AirBox-1 and AirBox-2, respectively. RMSE of Zhongming station decreases from 18.64 to 8.18.

For examining effects of spatial variation of landscapes on the bias of AirBox $PM_{2.5}$ measurements, the study constructed the calibration model with only one comparison site and the results show in Figure 5.8. The performance differences are small in each calibration results which the model used a single station data. It is intuitive to observe the best model performances were located along diagonal in Figure 5.8 that the low-cost sensor was calibrated by its own data. For the results, a calibration model with a single station data shows significant inconsistency reduction for AirBox and TWEPA measurements.

Moreover, the study examined the stability of TWEPA measurements interpolation

doi:10.6342/NTU202000011

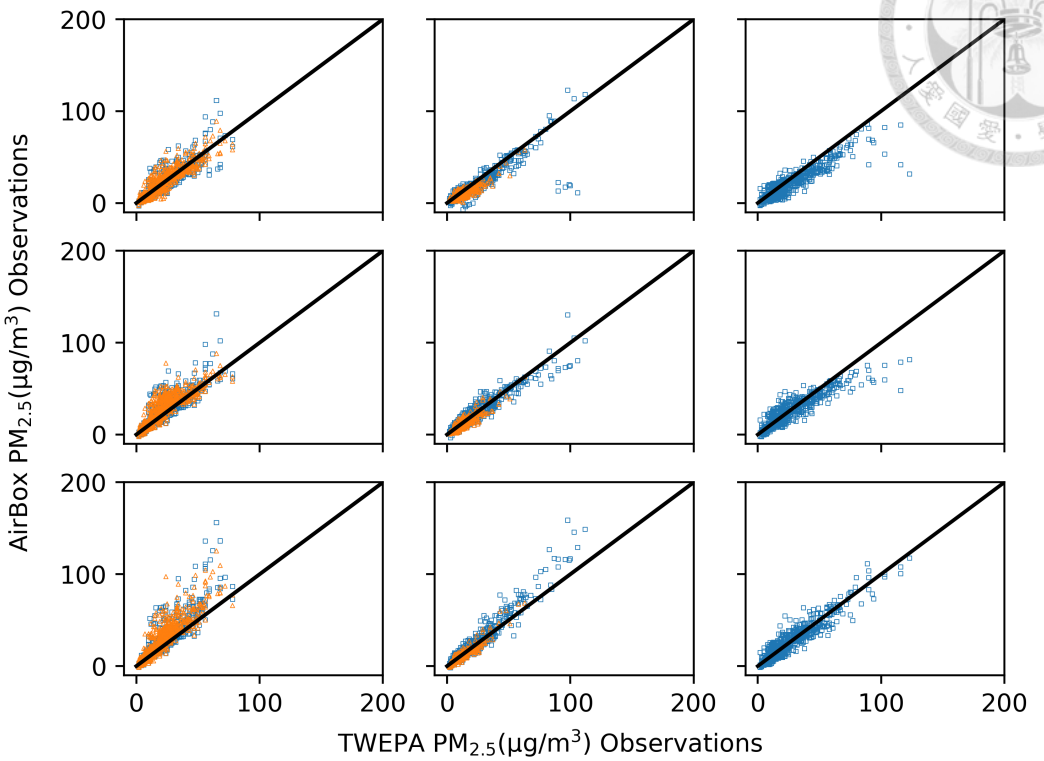


Figure 5.7: The results of applying a calibration model obtained from a speci c station on the calibration at the other stations. Left to right: Dali station, Shalu station, and Zhongming station. Top to bottom: the calibration model constructed by Dali, Shalu, and Zhongming observations, respectively.

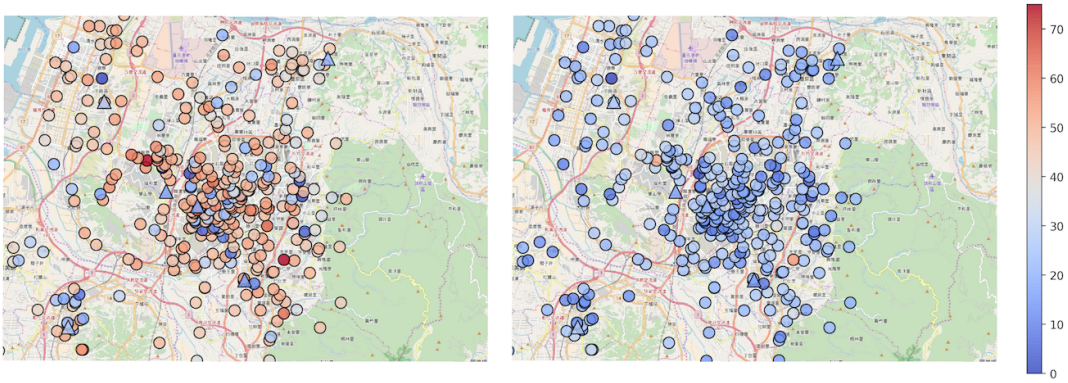


Figure 5.8: T The spatial evolution of the calibration process at 3 a.m. on Dec. 29th, 2017. Left: before calibration; right: after calibration. Triangular symbols are TWEPA stations and circular symbols are AirBox devices.

doi:10.6342/NTU202000011

that using estimated TWEPA measurements for the calibration. RMSE after calibration are 7.48, 9.31 at Dali station for AirBox-1 and AirBox-2, respectively. Shalu is 5.19 and 5.27 for AirBox-1 and AirBox-2, respectively. Zhongming is 6.56. The results shows the performance is close to using observed TWEPA measurements in the study.

5.6 Discussions

This study proposed an efficient data calibration approach for the $PM_{2.5}$ observations from low-cost sensors. In the emerging IoT era, the number of observations from all kinds of low-cost air quality sensors has been increased exponentially. To our knowledge, this study is one of the very first studies to systematically investigate the space-time data calibration approach for the low-cost $PM_{2.5}$ sensors. It has been known that the data quality of low-cost air quality sensors can varies with environmental conditions; however, it is still not very clear how these conditions can affect the uncertainties of observations, though these can also be highly associated with the quality of sensor itself. We proposed a GAM approach to empirically assess how the environmental conditions are associated with the observation biases in low-cost sensors, and how these associations could change across space and time. In previous analyses, the data calibration may commonly requires space-time homogeneity of environmental conditions; however, the suitability of homogeneous assumption could not easily be validated. In our proposed approach, the observations and comparisons of these associations can be an important reference to the quality and stability of the bias adjustment process for the IoT data.

In addition, to implement an operational data calibration approach, the computational cost is always a practical factor to be considered. Large amount of $PM_{2.5}$ observations would be reported from low-cost sensors and need to be calibrated in a relatively high frequency. It is required to assess the data calibration relationship frequently because the model performance can always not only be associated with observed but also unobserved environmental conditions. As a result, this study applied a GAM formulation because of its advantage of flexibility and computational efficiency. This approach allows us to quickly assess the model performance by by expert's assumptions or domain knowledge.

doi:10.6342/NTU202000011

In this study, the average time consumption of one-day calibration for all 438 sensors is about 21 seconds. (The computation was running at a laptop with 2.3 GHz Intel Core i5 processor.) Another consideration for operational data calibration approach is the data availability, in this study, the covariates used for the bias correction relationship mostly can directly obtained from the low-cost sensors in real-time. These covariates can be uncertain due to the limitations of these instruments; therefore, it can be important to check and update the calibration relationship dynamically for the operational purposes. The proposed calibration approach can explain the variability of the biases of $PM_{2.5}$ observations from the low-cost sensors with R-square of about 0.76 which is better than the results reported in previous studies [55].

Though the meteorological conditions have been considered to be an important factor to affect low-cost sensor data quality, several previous studies observed the $PM_{2.5}$ variation can relatively less associated with meteorological factors [51, 54], i.e., temperature, humidity, and ambient light. This study found that temperature has the nonlinear associations in which the observed biases in AirBox were mostly insignificant between 17^oC and 27^oC during the study area, and high associations while the temperature is out of this range. This finding is compatible with previous results [51]. The relative humidity shows its insignificance to the observed bias variation; however, to consider its nonlinearity, some positive association was shown while the relative humidity was high, particular for the range over 85%. However, it should be noted that all of the meteorological factors in our analysis were also measured by low-cost sensors, because they are readily available for every AirBox across space and time. Though the relatively high correlation, i.e., about 0.95 for temperature and 0.9 for relative humidity, the comparison presents the high inconsistency in observed levels between the relative humidity observation at the collocated AirBox and TWEPA stations. Furthermore, most of the relative humidity observations were as 100% which were not shown in TWEPA stations. In other words, in low-cost sensors, high uncertainty and bias can exist in both air quality and meteorological observations.

It has been known the space-time variation of $PM_{2.5}$ in Taichung has strong seasonality

doi:10.6342/NTU202000011

with which the $PM_{2.5}$ concentration is generally much higher during winter and spring than summer and autumn. This variation can mostly depend upon the change of meteorological conditions and human activities over time. In other words, the empirical relationship with different covariates in data calibration can possibly change accordingly across space and time. Though the time-varying nature of empirical relationships, it is important to maintain the stability of empirical relationships to mitigate the influence from data variance and uncertainty. Based upon this idea, the study used a 30-days window to allocate the data of $PM_{2.5}$ and their covariates for the dynamic assessment of empirical relationship for data calibration. In our case, this choice can not only guarantee the data abundance for the nonlinear relationship analysis but also maintain the observation ranges in $PM_{2.5}$ and covariates in similar ranges in adjacent days. Similarly, the empirical relationship for bias adjustment can also vary across the space. This study performed the cross-validation to investigate the spatial homogeneity of the calibration model. The spatial homogeneity is important for the application of data calibration on low-cost sensors located in the study area. In addition, sensor aging is also a potentially important factor for data calibration due to the degradation of the low-cost sensors. In our analysis, this aging factor was not shown since the data calibration approach was only conducted for two months. Furthermore, the information of sensor ages is not available in our dataset. In addition, the data calibration model can also change with respect to different manufacturers or installed chips. In other words, the data calibration result in this analysis is not applicable to the other low-cost $PM_{2.5}$ sensors.

Though the data calibration is crucial to obtain a reasonable $PM_{2.5}$ levels from low-cost sensors, in our study, the comparison of raw $PM_{2.5}$ measurements between low-cost sensors and regulatory stations showed relatively high correlation, i.e., Pearson's r ranging between 0.86 and 0.92 at all comparison sites. It implies the low-cost sensors in our study have capability of capturing the spatiotemporal trend of $PM_{2.5}$ variation without calibration. In other words, the raw low-cost observations can already be used to identify the space-time pollution hotspot to alert the public to the high $PM_{2.5}$ events or areas. The purpose of data calibration is to deliver a post-processing measurements that ensure high

doi:10.6342/NTU202000011

level of data quality from the low-cost sensors. Because the increasing awareness of adverse health impact of $PM_{2.5}$, the bias of the $PM_{2.5}$ observations from low-cost sensors can possibly induce the bias of the risk perception from the public. Even though the data accuracy can be significantly improved through the calibration, the high uncertainty can still be present in post-processed data. For example, as shown in Figure 5.7, the collocated raw and calibrated observations from low-cost sensors may not be the same. This study shows that, though the relatively better performance in our analysis, the averaged standard deviation of the calibrated $PM_{2.5}$ uncertainty is generally about 13.85% with respect to its adjusted levels. In other words, the bias-adjusted result and its associated uncertainty should be both considered in the assessment of space-time distribution of $PM_{2.5}$ levels, particularly for the purposes of risk communication.

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

Chapter 6

A High Performance Spatiotemporal Data Fusion Approach for Integrating $PM_{2.5}$ Hard and Soft Measurements

(*Unpublished*)

"Every word you say. I put my faith, and try to make it sense. Every move you take. I tell myself, it really makes sense."

法蘭黛樂團 **Frandé**《**Every Word**》

While low-cost sensors have been widely applied and deployed, reliability is not the only issue need to be concerned. How to integrate the measurements of low-cost sensors with regulatory stations is another serious issue to both government and citizen, because of the measuring method between low-cost sensors and regulatory stations is complete different. Generally, due to the high uncertainty, government agencies have regarded low-cost sensors as untrustworthy devices and the measurements are not suitable for reporting official air quality to public. This issue is also related to visualization of air quality monitoring that it is the most common way to show all air quality sources on the same map. When citizen has reached the mixed information, they usually could not distinguish between the

doi:10.6342/NTU202000011

reliability of different kinds measurements. Citizen would not notice effects of uncertainty of appearing numbers on the air quality map. Therefore, it is nature to draw a comparison between all numbers.

On the other hand, from the viewpoint of *Big Data*, more data means the potential of better problem addressing than before. Any kinds of data should be included into the analysis for improving the results. However, lack of realization from IT fields in environmental exposure assessment caused that they neglected the huge influence of uncertainty on social and political responsibility. Although calibration processes of low-cost sensors could be carried out afterwards, uncertainty of low-cost sensors still stands for the particular limitations than regulatory measurements.

At the present, based on the above reasons, it is urgent to develop a integrating framework with considering different reliability of regulatory stations and low-cost sensors measurements. In this chapter, we will used the same dataset from Chapter 4 that $PM_{2.5}$ measurements of low-cost sensors have been calibrated, and develop a data fusion framework base on Bayesian Maximum Entropy to integrate both regulatory stations and low-cost sensors.

6.1 The issue of high and low uncertainty air quality measurements in visualization and interpretation

Recently, the uptrend in the development of IoT industry has brought great benefit to air quality monitoring. One of the many benefits of IoT-based air quality devices is to lower the entry level threshold for air quality monitoring. With these commercial sensors, citizens are engaged to collect and share air quality information due to its low cost and mobility. Furthermore, it substantially enhanced the resolution of air quality information in space and time, namely, increased the volume of spatiotemporal air quality data. The large amounts of air quality information can be used not only for mapping air pollution, identifying the sources, and tracking changes but, in further, predict extreme air quality events. Low-cost sensors certainly could be used to raise public awareness about local

doi:10.6342/NTU202000011

air pollution that air pollutant concentrations stand for the levels of adverse health effects people exposed to[13]. Since the performance is one of the most concern when low-cost sensors are applied for the task of monitoring, a variety of calibration works have been carried out along with sensor system deployments. Unlike the regulatory air quality monitoring stations with certified reference instruments, U.S. EPA has indicated that precision and bias are the most concerns of low-cost air sensors [57]. Either field or laboratory calibration have found that the measurements of low-cost sensors would vary because of the changes of external environment[10, 30, 24, 70, 71, 63, 13]. Therefore, it would be inevitable to face the uncertainty problem in the usage of low-cost sensor measurements.

This uncertainty problem has strongly associated with exposure assessment in risk analysis, moreover, the variation of air quality measurements will seriously affect public risk awareness or perception. Even though the uncertainty of low-cost sensors could be quantified through scientific researches, misunderstandings in citizens and communities are commonly caused by lack of uncertainty information. In the era of *Big Data*, data visualization tools and techniques have played an important role in efficiently displaying comprehensive data characteristics. From the view of government agencies, providing air quality information is an issue related to responsibility that embodies governmental accountability. The websites of air quality maps built by governments are collected the measurements from legislative reference or equivalent stations, e.g., AirNow[95], European Air Quality Index[31], and Taiwan Air Quality Monitoring Network[88]. Governmental agencies clearly understand air quality information is linked to public health concerns that inexplicit measurements would produce negative impacts on society. On the other hand, environmental or IT communities are conventionally aggregate all types of air quality monitoring measurements that display on the same map[36]. Besides, people are not familiar to interpret air quality information with another variance map(uncertainty map). Hence, putting both certain and uncertain air quality measurements together without further the explanation of uncertainty is a very critical problem for risk perceptions and communications.

To integrate both certain and uncertain air quality measurements such as regulatory sta-

doi:10.6342/NTU202000011

tions and low-cost sensors, respectively. Data fusion technique is widely used to combine multiple data sources that produce a more consistent and accurate results. Generally speaking, geostatistics can be seen as one data fusion technique that focuses on incorporating different spatial or spatiotemporal datasets. Geostatistics has another essential feature that it could estimate unmonitored locations based on the observations. The most famous geostatistical algorithm is Gaussian process regression as known as Kriging method[60, 23]. There is one study has applied Kriging method with Optimal Linear Data Fusion theory to map a finer resolution of spatiotemporal $PM_{2.5}$ concentrations[66]. Because of low-cost sensors are considered with high uncertainty, it is a problematic situation to integrate certain and uncertain data together in using data fusion techniques. At the beginning, Kriging method was not able to account for uncertain information. There were studies proposed transformed Kriging methods that could include uncertain measurements into the interpolation[56, 90]. However, theses methods required complex statistical inference or assumptions with great effort and would be inefficient on large datasets.

In the present study, we will use a novel geostatistical approach - Bayesian Maximum Entropy(BME) method. BME method is a spatiotemporal geostatistical approach based upon epistemic framework. This knowledge based synthesis framework that collect knowledge sources relevant to the interest[19]. The previous study has showed that BME method has more flexible integrating framework and better interpolation results[20]. In addition, large amount of low-cost air quality sensors caused an apparent problem for the computational cost of data fusion process. Due to expected fast and voluminous air quality data, this study modifies the inner integrating algorithm of BME method to develop a high performance data fusion framework.

6.2 Deployment and calibration of commercial $PM_{2.5}$ sensors in Taiwan

Currently, there are seventy-six air quality monitoring stations operated by Taiwan Environmental Protection Administration (TWEPA) throughout the island. These stations

doi:10.6342/NTU202000011

constantly collect six criteria pollutants including sulfur dioxide (SO_2), nitrogen oxides (NO_x), ozone (O_3), carbon monoxide (CO), PM_{10} (particles less than 2.5 micrometers in diameter) and $PM_{2.5}$ (particles less than 10 micrometers in diameter). Among them, there are five stations located within Taichung metropolitan area, which is the second largest city in Taiwan with the population size of 2.79 million in 2018. Recently, awareness of $PM_{2.5}$ health issues in Taichung is rapidly raising because the largest coal-fired power plant in the world is located in the city. This study used the hourly $PM_{2.5}$ from the TWEPA stations in Taichung.

Since the early 2016, the popularity of low-cost air quality sensors has been rapidly increased island-wide, particularly in metropolitan areas. Over 5,000 air quality sensors have been installed in the last two years in Taiwan. In Taichung, local communities have volunteered to install low-cost $PM_{2.5}$ sensors of various types over the most of highly urbanized parts of the study area. Among these low-cost sensors, the majority belongs to the AirBox air quality detection system produced by Edimax Inc[28], that currently account for more than 85% of the all recorded low-cost sensors. The popularity of this particular sensor primarily resulted from its high accessibility and its manufacturer promoting strategy. To have the same variability and stability of sensing quality, this study only considers $PM_{2.5}$ observations from AirBox in our analysis.

Before carry out data fusion technique, the measurements of low-cost sensors firstly are calibrated by the approach developed in the previous study[63]. The calibration process is based on generalized additive model which is further applied to AirBox. The study performed a field calibration that collecting both measurements of low-cost sensors and the regulatory stations, and investigated the space/time bias between the low-cost sensors and regulatory stations. Calibration results showed that the calibration approach could explain the variability of the biases from the low-cost sensors with R-square of 0.76. In addition, the calibration model could quantify the uncertainty of the low-cost sensors observations and the average standard deviation is about 13.85% with respect to its adjusted levels.

To demonstrate the effects of multiple data sources for mapping, the study selects an

doi:10.6342/NTU202000011

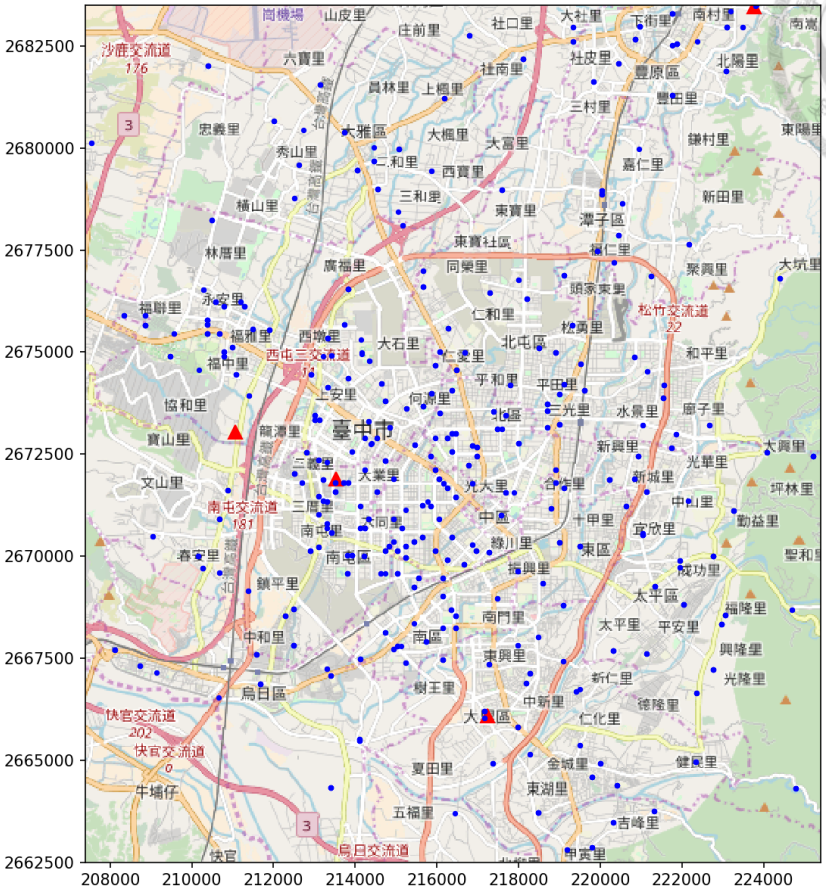


Figure 6.1: Spatial distribution of TWEPA regulatory stations (red triangles) and AirBox devices (blue circles) in the selected Taichung metropolitan area.

area with specific spatial range (Easting: from 207,400.0 to 225,400.0; Northing: from 2,662,500.0 to 2,683,500.0. EPSG: 3826 Projected coordinate system for Taiwan) There are total 4 TWEPA regulatory stations and 1034 AirBox devices are within the area. The spatial distribution of AirBox sensors and TWEPA stations in the selected area is shown in Figure 6.1.

doi:10.6342/NTU202000011

6.3 Data fusion algorithm - BME method

BME method is a knowledge-based data fusion geostatistical algorithm that the knowledge base $\mathcal{K}$ is constructed by two prime knowledge bases: 1.) the *general* knowledge base $\mathcal{G}$: includes all knowledge bases of natural characteristics, which can be scientific laws, empirical relationships, and theoretical space-time dependence models; and 2.) the *specificatory* knowledge base $\mathcal{S}$: includes all knowedge bases that are specific to the region of interest and further split up into: *hard* data (i.e. exhibiting a satisfactory level of accuracy), and *soft* data that represent uncertainty in the observations (including secondary information, imperfect observations, categorical data, and fuzzy inputs). On the other hand, $\mathcal{K} = \mathcal{G} \cup \mathcal{S}$ represents the total available knowledge.

Considering the space-time distribution of $PM_{2.5}$ as an spatiotemporal random field (S/TRF). The S/TRF can be viewed as a collection of field realizations associated with the correlated random variables in space/time. Assume that $\boldsymbol{\chi_{map}} = (\chi_1, \chi_2, ..., \chi_m)$ are the space/time random variables at the space/time locations $\boldsymbol{p_{map}} = (p_1, p_2, ..., p_m)$ and $p = (s, t)$ is a space/time location that s and t denote spatial and temporal coordinates, respectively. A realizations of the map such as observed $PM_{2.5}$ concentrations at these locations can be denoted by the vector $\boldsymbol{x_{map}} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_m})$ The map $\boldsymbol{\chi_{map}}$ consists of observations $\boldsymbol{\chi_{data}}$ including *hard* data $\boldsymbol{\chi_h}$ and *soft* data $\boldsymbol{\chi_s}$, and *unobserved* values $\boldsymbol{\chi_k}$. The total knowledge synthesis in BME method is based on Bayesian inference. Operationally, the probability of a map $\boldsymbol{\chi_k}$ given the total knowledge $\mathcal{K} = \mathcal{G} \cup \mathcal{S}$ is following,

$$\begin{aligned} f_{\mathcal{K}}(\boldsymbol{\chi_k}) \equiv f_{\mathcal{K}(\boldsymbol{\chi_k}|\boldsymbol{\chi_h},\boldsymbol{\chi_s})} &= \frac{\frac{\partial}{\partial \boldsymbol{\chi_k}} F_{\mathcal{K}}(\boldsymbol{\chi_k}, \boldsymbol{\chi_h}, \boldsymbol{\chi_s})}{F_{\mathcal{K}}(\boldsymbol{\chi_h}, \boldsymbol{\chi_s})} \\ &= \frac{\int_{-\infty}^{\infty} f_{\mathcal{G}}(\boldsymbol{\chi_k}, \boldsymbol{\chi_h}, \boldsymbol{\chi_s}) \delta_S(\boldsymbol{\chi_h}) f_{\mathcal{S}}(\boldsymbol{\chi_s}) \mathrm{d}\boldsymbol{\chi_h} \mathrm{d}\boldsymbol{\chi_s}}{\int_{-\infty}^{\infty} f_{\mathcal{G}}(\boldsymbol{\chi_h}, \boldsymbol{\chi_s}) \delta_S(\boldsymbol{\chi_h}) f_{\mathcal{S}}(\boldsymbol{\chi_s}) \mathrm{d}\boldsymbol{\chi_h} \mathrm{d}\boldsymbol{\chi_s}} \end{aligned} \tag{6.1}$$

where $f_{\mathcal{G}}(\cdot)$ and $f_{\mathcal{S}}(\cdot)$ are the pdfs for the general knowledge and specificatory knowledge bases, respectively. $\delta(\cdot)$ is a function that expresses the assimilated hard data corresponds

doi:10.6342/NTU202000011

to the specificatory knowledge base as follows,

$$\delta(\boldsymbol{\chi_h}) = \begin{cases} \boldsymbol{\chi_h}, & \text{if } \boldsymbol{\chi_h} \in \mathcal{S}. \\ 1, & \text{otherwise.} \end{cases} \tag{6.2}$$

At the present study, *hard* data and *soft* data in this data fusion framework are the measurements of TWEPA stations and AirBox devices, respectively. From Eq. (6.1), the integral form is used to merge $\mathcal{G}$ and $\mathcal{S}$ including multiple $PM_{2.5}$ measurement sources in BME method. Due to the limited number of *hard* data (only four TWEPA stations in this case), the dimension of the integration is highly depending on the number of *soft* data used. Geostatistics conventionally assumes second-order stationarity, in other words, covariance function is used to describe the relationship between two points that is only dependent on the distances. This technique could efficiently reduce the volumes of data in the analysis since it simple exclude those measurements that too far away from the unobserved location. However, enormous low-cost sensors has brought a unimaginable spatial resolution. For example, if we include all *soft* information in space into the consideration, it will be an over one thousand dimensions integration. (See the example below. Eq. (6.3))

$$\int_{[-\infty,\infty]^n} f_{\mathcal{S}}(\boldsymbol{\chi_s}) \mathrm{d}\boldsymbol{\chi_s} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathcal{S}}(\chi_{s_1}, \ldots, \chi_{s_n}) \mathrm{d}\chi_{s_1}, \ldots, \chi_{s_n} \tag{6.3}$$

n: the numbers of *soft* data. It is easy to approximate the integral in low dimensions by a product rule. But, with high dimensional integration, computational cost of product rules cannot be afforded to evaluate the integrand.

6.4 High performance integration with Quasi-Monte Carlo method

Nowadays, with these numerous low-cost sensors, high dimensional integration problems has become an urgent issue in the BME data fusion framework. Namely, it has commonly

doi:10.6342/NTU202000011

suffered from the "*curse of dimensionality*" in the era of *Big Data*. For breaking the curse, Quasi-Monte Carlo (QMC) methods will be introduced and applied to BME method in this study.

Random sampling is broadly used to solve high dimensional problems. The well-known random sampling technique the Monte Carlo (MC) method is a very simple and widely used method. Let us consider an integral over d-dimensional unit hypercube.

$$I(f) = \int_{[0,1]^d} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{6.4}$$

The MC method approximates the integral, $Q_N(f)$, by generating and averaging N random samples points $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_N$ of the function, which are independent and uniformly distributed over the hypercube (Eq. (6.5)).

$$Q_N(f) = \frac{1}{N} \sum_{k=1}^{N} f(\boldsymbol{i}_k) \tag{6.5}$$

The MC method has the advantage of producing an unbiased estimate of the integral, i.e., $\mathbb{E}[Q_N(f)] = I(f)$, and its rate of convergence is $\mathcal{O}(1/\sqrt{N})$ that means it is often too slow for practical applications.

QMC methods are generally have the identical form as MC method but instead of randomly generate sample points, QMC methods choose random points deterministically to have a faster rate of convergence. Theoretically, QMC method can have $\mathcal{O}((\log N)^d/N)$ convergence rate, however, it is limited in practice with the error bound issue. Fortunately, modern QMC methods provide alternatives by working with *weighted function spaces*. In this study, *lattice rules* is used to achieve uniformity of sample points. For the details on QMC methods and lattice rules, please refer to [26].

6.5 Data fusion for $PM_{2.5}$ *hard* and *soft* measurements

For the purpose of examining the impact of multiple data sources in the visualization of $PM_{2.5}$ levels. BME method with QMC methods are used to estimate $PM_{2.5}$ concentra-

doi:10.6342/NTU202000011

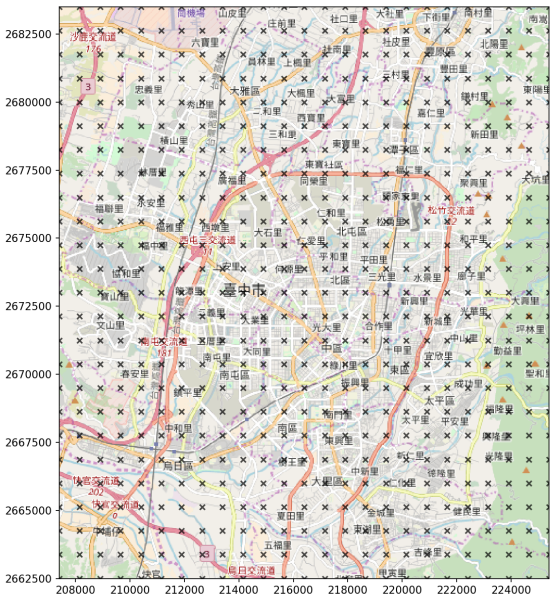


Figure 6.2: Spatial distribution of 625 estimated locations (black cross markers).

tions of the study area at a selected timestamp. There are total 625 spatial locations to be estimated (See Figure 6.2).

This study have designed five scenarios to compare effects of using $PM_{2.5}$ measurements with different levels of uncertainty. The study assumes the bias of AirBox measurements is normally distributed. There is analytical form if both $\mathcal{G}$ and $\mathcal{S}$ are Gaussian distributions under BME framework. Hence, for the purpose of revealing the capability of QMC methods in high dimensional integration, each *soft* data has transformed to Gaussian-like piece-wise linear distribution in Scenario 5. (The computation run at a laptop with 2.3 GHz Intel Core i5 processor.)

- **Scenario 1: only uses TWEPA measurements**

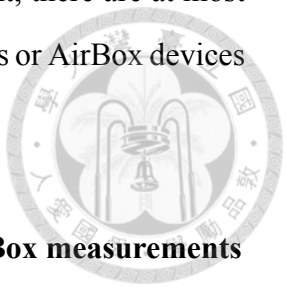
  At a estimation point, there are at most 5 nearest $PM_{2.5}$ measurements in space and time from TWEPA regulatory stations will be included.

- **Scenario 2: uses both TWEPA and AirBox measurements which is treated as hard data**

  The measurements of AirBox devices are treated as certain data, in other words, all

doi:10.6342/NTU202000011

measurements are belonging to *hard* data. At a estimation point, there are at most 30 nearest $PM_{2.5}$ measurements from TWEPA regulatory stations or AirBox devices in space and time will be included.



- **Scenario 3: uses both TWEPA and a small numbers of AirBox measurements**
  At a estimation point, there are at most 5 nearest $PM_{2.5}$ measurements from TWEPA regulatory stations and 5 nearest $PM_{2.5}$ measurements from AirBox devices in space and time will be included.

- **Scenario 4: uses both TWEPA and a large numbers of AirBox measurements**
  At a estimation point, there are at most 5 nearest $PM_{2.5}$ measurements from TWEPA regulatory stations and 50 nearest $PM_{2.5}$ measurements from AirBox devices in space and time will be included.

- **Scenario 5: uses both TWEPA and a large numbers of AirBox measurements**
  At a estimation point, there are at most 5 nearest $PM_{2.5}$ measurements from TWEPA regulatory stations and 25 nearest $PM_{2.5}$ measurements from AirBox devices in space and time will be included. *soft* data is transformed into Gaussian-like piece-wise linear probability density distribution.

The spatiotemporal estimation in all scenarios above will use the same theoretical space-time nested covariance model c_X as below

$$c_X(h,\tau) = c_1 \exp(-3\frac{a_\tau h + a_h \tau}{a_{h_1} a_{\tau_1}}) + c_2 \exp(-3\frac{a_\tau h + a_h \tau}{a_{h_2} a_{\tau_2}}) \tag{6.6}$$

where $h = \mid s_i - s_j \mid$ and $\tau = t_i - t_j$ denote the spatial and temporal lags, respectively, between any pair of space-time points $p_i = (s_i, t_i)$ and $p_j = (s_j, t_j)$. c_1 and c_2 are the sill coefficients. a_{h_i} and a_{τ_i} are the spatial and temporal autocorrelation ranges that characterize the different covariance models nested in Eq. (6.6), respectively. The empirical and theoretical covariance model are as shown in Figure 6.3.

doi:10.6342/NTU202000011

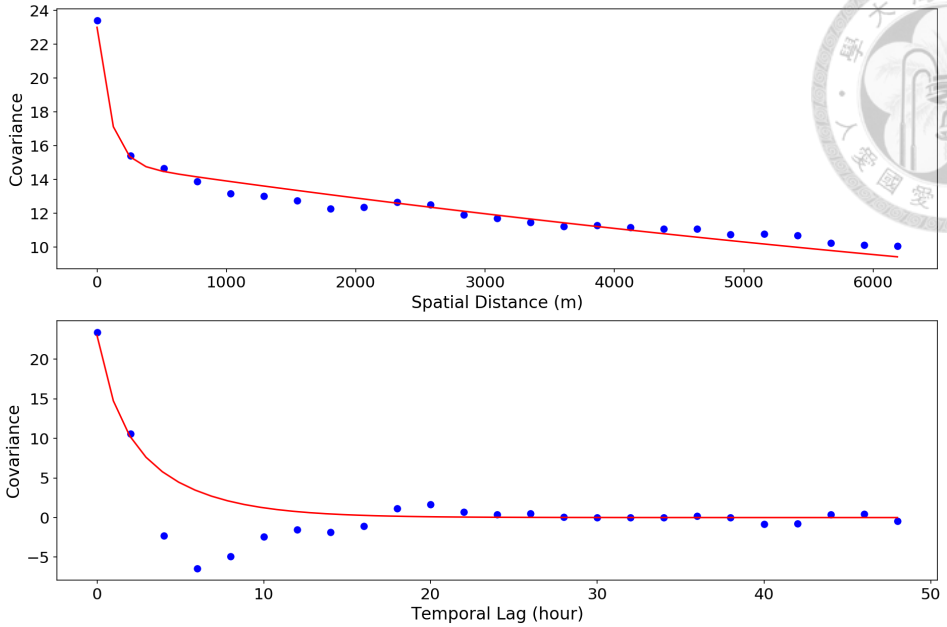


Figure 6.3: The empirical and theoretical covariance model (Above: spatial; Below: temporal). Blue dots are empirical covariances that calculated from *hard* data and mean values of *soft* data. Red line is the fitted theoretical covariances.

6.6 The evolution of $PM_{2.5}$ levels mapping

The study performed a efficient data fusion framework of multiple $PM_{2.5}$ measurements sources that accounted for the different levels of uncertainty. Effects of certain and uncertain measurements are examined through five designed scenarios. The 4 AM December 23, 2017 was selected as the temporal estimation point. In Scenario 1, only the measurements of TWEPA regulatory stations were considered in the interpolation. The result shows unusual distribution of $PM_{2.5}$ concentrations and the estimation which has high variance as shown in Figure 6.4. It is no surprise that $PM_{2.5}$ information is so scarce for unobserved locations from nearby measurements. The result of Scenario 2 is opposite to Scenario 1 that $PM_{2.5}$ distribution look extremely smoothing with lower variance as shown in Figure 6.5. Because Scenario 2 regarded the measurements of AirBox as certain data, available $PM_{2.5}$ information significantly increased for unobserved locations.

Scenario 3 and Scenario 4 used both *hard* (TWEPA) and *soft* (AirBox) information of $PM_{2.5}$. The difference between theses two scenarios is the numbers of *soft* data used in the

doi:10.6342/NTU202000011

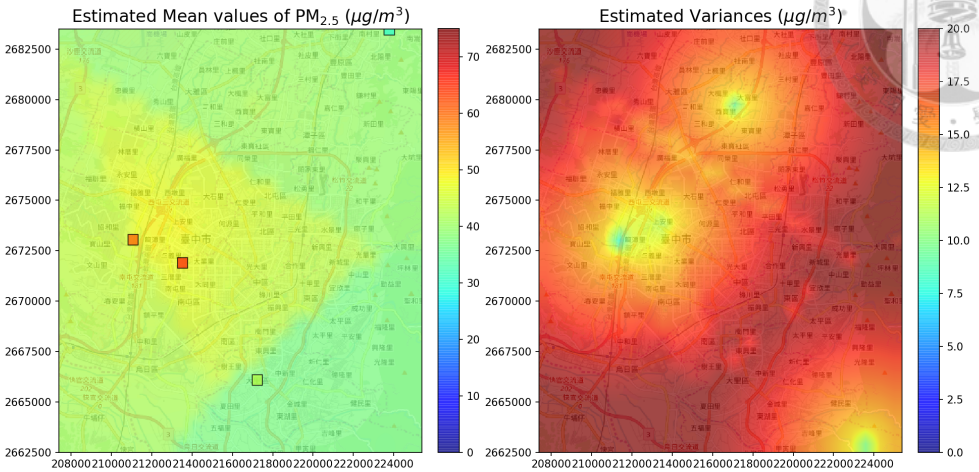


Figure 6.4: Scenario 1 estimation of spatiotemporal $PM_{2.5}$ concentrations. *Left*: expected values of estimation; *Right*: variances of estimation.

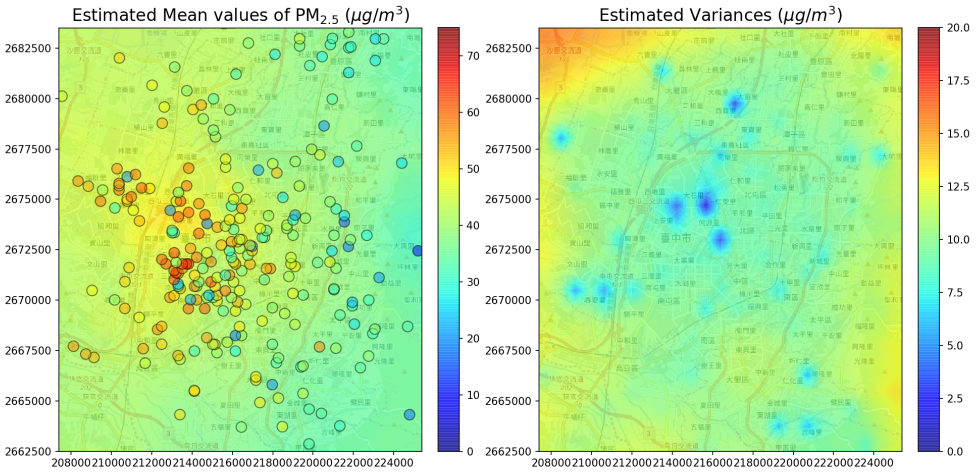


Figure 6.5: Scenario 2 estimation of spatiotemporal $PM_{2.5}$ concentrations. *Left*: expected values of estimation; *Right*: variances of estimation.

doi:10.6342/NTU202000011

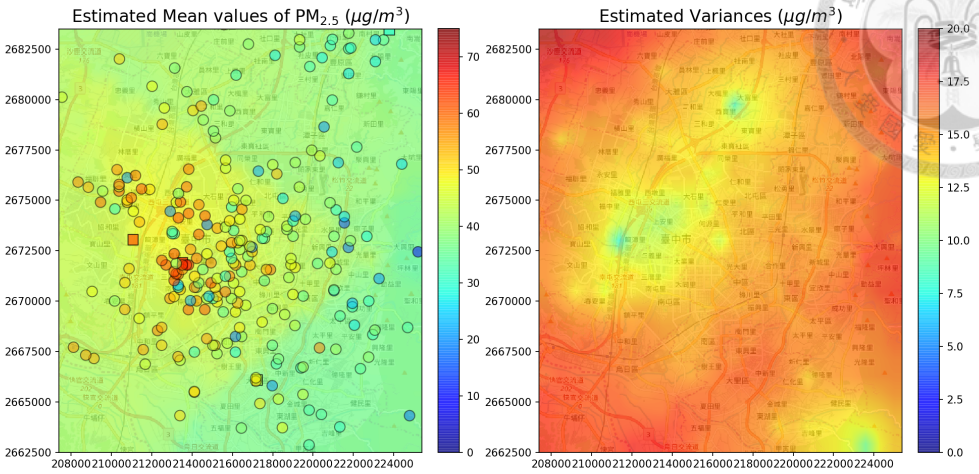


Figure 6.6: Scenario 3 estimation of spatiotemporal $PM_{2.5}$ concentrations. *Left*: expected values of estimation; *Right*: variances of estimation.

estimation. Figure 4.6 and Figure 6.7 are the estimation results of Scenario 3 and Scenario 4, respectively. From the results of Scenario 3, it is clear to see a synthetic boundary of $PM_{2.5}$ concentrations on the map which is similar to the result of Scenario 1. However, Scenario 3 has lower variance than Scenario 1. Scenario 4 which took account of more AirBox measurements displays a reasonable $PM_{2.5}$ distribution that there is no obvious abnormality in the distribution in comparison to Scenario 3.

Figure 6.8 shows the estimation result of Scenario 5. The result is noticeably differ from other scenarios that the expectation of $PM_{2.5}$ concentrations has shown more details of spatial variation. The variance map in Figure 6.8 exhibits there are many clusters of high variance existing.

6.7 Discussions

The emergence of IoT sensors has made a great contribution to environmental monitoring due to their low complexity and cost. The development of low-cost sensors apparent

doi:10.6342/NTU202000011

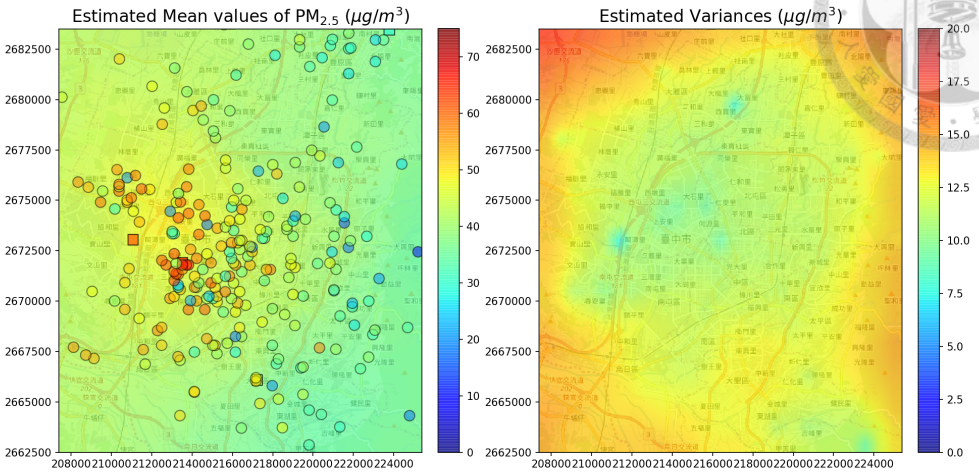


Figure 6.7: Scenario 4 estimation of spatiotemporal $PM_{2.5}$ concentrations. *Left*: expected values of estimation; *Right*: variances of estimation.

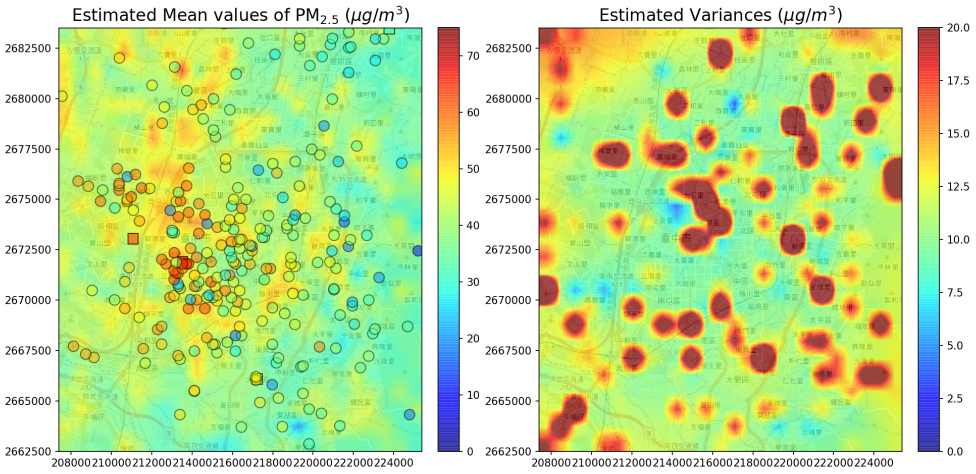


Figure 6.8: Scenario 5 estimation of spatiotemporal $PM_{2.5}$ concentrations. *Left*: expected values of estimation; *Right*: variances of estimation.

doi:10.6342/NTU202000011

become a solution of increasing the spatial and temporal resolution for environmental monitoring. The drawback of low-cost sensors is that the measuring method used is different from regulatory instruments which are required to meet legislative criteria, so the measurements of low-cost sensors should be considered with uncertainty. For example, instead of measuring particulate mass directly such as regulatory instruments, low-cost PM sensors use light-scattering method to count particles that pass through the optical cell. There must have concerns about precision and bias of low-cost sensors. Because it is an ongoing technology that environmental data analysis is still not matured, low-cost sensor measurements are commonly put together with regulatory station measurements on visualization, in further, causes misleading of risk perception.

This study performed a spatiotemporal data fusion application of $PM_{2.5}$ *hard* and *soft* data and carried out a series of estimation under various scenarios. The deployment of low-cost $PM_{2.5}$ sensors definitely have the positive impact on providing a finer resolution of $PM_{2.5}$information from the comparison between Scenario 1 and Scenario 2 that more certain data would generate more smoothing distribution with low variances. However, if the visualization neglects to consider the uncertainty of *soft* data which is an easier and accepted way within communities, this would result in influence on risk perception and communication of adverse health. For example, the expectation of $PM_{2.5}$ levels in both Scenario 2 (Figure 6.5) and Scenario 4 (Figure 6.7) is similar but the map of variances has significant dissimilarity. The estimation of Scenario 2 only considered the uncertainty which caused by the distances between estimated and data points. In contrast, the estimation of Scenario 4 not only consider the uncertainty due to the distances but the uncertainty of *soft* data. From the variance map of Figure 6.7, it shows the estimation of $PM_{2.5}$ levels could vary by one order of magnitude in particular area which means the levels of risk could be total different.

Besides, the study designed a special case, Scenario 5, to examine the QMC methods under the operational BME framework. With the powerful high dimensional integration QMC methods, BME framework are not limited to the amounts and Gaussian assumption of *soft* data in practical as the result show in Figure 6.8. It can be expected that there will

doi:10.6342/NTU202000011

be more and more low-cost sensors developing and deploying, moreover, the bias of each low-cost sensor is not necessarily normally distributed in reality. Hence, the result of Scenario 5 has proved this data fusion framework is well prepared for the future applications. The design of Scenario 4 and 5 were expected to have similar $PM_{2.5}$ distribution, however, the result of Scenario 5 generated a unreasonable state. Because we used piece-wise linear function to approximate Gaussian distribution of soft data, upper bound and lower bound of approximated pdf would be cut out. Hence, errors of numerical integration might be the cause of difference between Scenario 4 and 5.

By conducting these scenarios, the present study wants to emphasize the importance of the uncertainty realization on visualization in environmental data analysis. A environmentalist should be aware of the effects of data uncertainty in this new data era while produces any types of environmental information.

doi:10.6342/NTU202000011

doi:10.6342/NTU202000011

Chapter 7

To the End of the Journey

”我們的雙手緊握，在黑暗中，我
不會把你放開，小心翼翼的走著，
再過不久，一定就能看見光。”

鄭宜農 **Enno Cheng**《光》

The main theme of the dissertation is not about what new algorithms or scientific discoveries I have accomplished during my doctoral studies. It is more about my personal experiences through grant proposals writing, projects execution, conferences participation, and journal articles reading etc. in these years. As an environmentalist, it is great to know the society concerns about the land where they live more than ever. As an environmental scientist, it is honored that make efforts to understand more about the environment. As an environmental data scientist, it is happy to see environmental information is easier to be reached. In this present era, data, algorithms, and hardwares have made tremendous success in a variety of applications and ignites the spark of imagination for building the new future. Honestly, new technology looks very fancy and attractive to environmental sciences. There are a lot of imagination to apply new techniques to environmental study and lead to new findings. However, in my opinion, environmental information is strongly associated with social responsibility. It should be taking careful steps to these applications.

To my knowledge, the most important issue is what kinds and meanings of environmental information people received that also means the works in risk communication. In

doi:10.6342/NTU202000011

reality, no one would care about environmental quality only if there is risk that bad environment would cause adverse health effects. Hence, environmental information somehow can affect people's risk perception, in further, becomes a political leverage. Environmental scientists are responsible for helping out others no matter government agencies or public to realize the representation of environmental information with their expertise. In my experience, people from IT field believes that the devices they developed have promoted environmental monitoring to the next level. But they are not in charge of explanation about the numbers because they thought citizens should be able to distinguish the uncertainty from devices by themselves. The truth is citizens either do not care or lack of knowledge about accuracy, precision and uncertainty. This phenomenon will cause cognitive gap between perception and sensation, in further, lead to biased decision-making.

There are three studies are used to convey the core concept of this dissertation. Firstly, we have built an early warning system for Dengue fever in the southern Taiwan. In recent years, a large amount of data that cannot be reached by public before are released along with the concept of *Open Government*. Hence, environmental data analysis would gain many applying opportunities to integrate multiple data sources with cross-disciplinary knowledge nowadays. In the EWS study, we used the relationship between meteorological factors and incidences considering time lagged effects to predict Dengue fever. The model successfully allowed us to predict the outbreak one week ahead before happening. Because disease transmission is a complex process, at the time, we used stochastic modeling to include all information that we did not have enough information. The results showed that meteorological factors only accounted for a small portion of Dengue fever spreading in the study area. It also means that if we have other information related to Dengue fever, we can explain less with stochastic modeling. After several years, now we can access more information such like Dengue fever prevention and control strategies and dominant serotype in each year. These information are expected to give us better understanding about the transmission and significantly improve the early warning system in the future. Secondly, we have developed an efficient calibration model for low-cost $PM_{2.5}$ sensors. The fast growth of IoT industry has shown great improvement in environmental

doi:10.6342/NTU202000011

monitoring but the emergence of low-cost sensors has brought new concerns as well. The study would like to point out the importance of the uncertainty of new IoT sensors, especially, these measurements could be connected to adverse health impacts. The calibration model was based on the bias relationship between low-cost $PM_{2.5}$ sensors and regulatory stations. Besides, the low-cost sensors were dynamically calibrated everyday to reduce influences of sensor aging or dust accumulations. However, the calibration model only included very few variables due to limited measuring factors from low-cost sensors and $PM_{2.5}$ is considered to be highly dependent on weather condition. The bias relationship in the study might not be suitable to other places even are the same low-cost sensors. But it is worth to analyze the measurements in other cities to see if the bias relationship has spatial dependence or not. Lastly, BME method was used to integrate multiple data sources, specifically, data with different levels of uncertainty. The study illustrated that the evolution of $PM_{2.5}$ mapping under different circumstances, e.g., inspect the volume of *hard* and *soft* data used in estimation. Today, it is a crucial issue for environmentalists to demonstrate the necessity of considering data uncertainty in the visualization. The different ways of environmental information visualization would have a huge effect on people's risk perception. This is easy to be neglected by who is lack of knowledge of risk. The different results of Scenarios gave an interesting research topic that it will be very valuable to assess benefits mapping within interpolation methods and data used.

The purpose of this dissertation is not encouraging to abandon all these new technologies. Here is to emphasize the consequence of environmental information publicly releasing. The main point is close to Environmental Information Disclosure program which was proposed by The World Bank[91]. Every character has their own viewpoint about environmental information and everyone's involvement and disagreement often turn out a conflict of interest. Therefore, using the prospective in risk communication is the key to have social responsibility for ensuring that the balance between environmental information providing and receiving. In addition, lacking of physical laws inside is one of the most concerns while interpret the results from data analysis. Like all kinds of machine learning algorithms are based on statistical relationships between input and output vari-

doi:10.6342/NTU202000011

ables. Right now, these relationships still not strong enough to be the evidence in risk assessment. Thus, about the future of environmental data analysis, one question is worth thinking that how to integrate physical properties into data analysis to boost its persuasion power.

In conclusion, the dissertation stands at the perspective of risk analysis to inspect what kind of role that environmental data sciences have played. Environmental data analysis have taken an very important position in this new data ear. However, the main theme of the dissertation would like to provoke a idea of *responsibility* that environmental data scientists not only **analyze the information** but **be responsible for the information**. The dissertation seeks to open the way for environmental data analysis which is associated with risk management, in further, possible contributions to environmental protection.

doi:10.6342/NTU202000011

Bibliography

[1] A. R. Al-Ali, I. Zualkernan, and F. Aloul, *A mobile GPRS-sensors array for air pollution monitoring*, IEEE Sens. J., 10 (2010), pp. 1666–1671.

[2] J. Angulo, H. L. Yu, A. Langousis, A. Kolovos, J. Wang, A. E. Madrid, and G. Christakos, *Spatiotemporal Infectious Disease Modeling: A BME-SIR Approach*, PLoS One, 8 (2013).

[3] J. M. Angulo, H.-L. Yu, A. Langousis, A. E. Madrid, and G. Christakos, *Modeling of space–time infectious disease spread under conditions of uncertainty*, 2012.

[4] Array of Things, *Array of Things*.

[5] N. S. Asmarian, A. Ruzitalab, K. Amir, S. Masoud, and B. Mahaki, *Area-to-Area Poisson Kriging analysis of mapping of county- level esophageal cancer incidence rates in Iran.*, Asian Pac. J. Cancer Prev., 14 (2013), pp. 11–3.

[6] S. Athey and G. W. Imbens, *The state of applied econometrics: Causality and policy evaluation*, in J. Econ. Perspect., 2017.

[7] F. Baker, *Risk communication about environmental hazards*, J. Public Health Policy, (1990).

[8] P. Bhatia, *Introduction to Data Mining*, in Data Min. Data Warehous., 2019.

[9] C. R. Bilder and J. M. Tebbs, *Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs.*, Biom. J., 47 (2005), pp. 502–516.

doi:10.6342/NTU202000011

[10] D. M. Broday, A. Arpaci, A. Bartonova, N. Castell-Balaguer, T. Cole-Hunter, F. R. Dauge, B. Fishbain, R. L. Jones, K. Galea, M. Jovasevic-Stojanovic, D. Kocman, T. Martinez-Iñiguez, M. Nieuwenhuijsen, J. Robinson, V. Svecova, and P. Thai, *Wireless distributed environmental sensor networks for air pollution measurement-the promise and the current reality*, Sensors (Switzerland), 17 (2017).

[11] D. V. Canyon, J. L. Hii, and R. Müller, *Adaptation of Aedes aegypti (Diptera: Culicidae) oviposition behavior in response to humidity and diet.*, J. Insect Physiol., 45 (1999), pp. 959–964.

[12] O. D. Cardona, M. K. Van Aalst, J. Birkmann, M. Fordham, G. Mc Gregor, P. Rosa, R. S. Pulwarty, E. L. F. Schipper, B. T. Sinh, H. Décamps, M. Keim, I. Davis, K. L. Ebi, A. Lavell, R. Mechler, V. Murray, M. Pelling, J. Pohl, A. O. Smith, and F. Thomalla, *Determinants of risk: Exposure and vulnerability*, in Manag. Risks Extrem. Events Disasters to Adv. Clim. Chang. Adapt. Spec. Rep. Intergov. Panel Clim. Chang., 2012.

[13] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova, *Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?*, Environ. Int., (2017).

[14] L. J. Chen, Y. H. Ho, H. C. Lee, H. C. Wu, H. M. Liu, H. H. Hsieh, Y. T. Huang, and S. C. C. Lung, *An Open Framework for Participatory PM2.5 Monitoring in Smart Cities*, IEEE Access, 5 (2017), pp. 14441–14454.

[15] M.-J. J. Chen, C.-Y. Y. Lin, Y.-T. T. Wu, P.-C. C. Wu, S.-C. C. Lung, and H.-J. J. Su, *Effects of extreme precipitation to the distribution of infectious diseases in Taiwan, 1994-2008*, PLoS One, 7 (2012), p. e34651.

[16] S. C. Chen and M. H. Hsieh, *Modeling the transmission dynamics of dengue fever: Implications of temperature effects*, Sci. Total Environ., 431 (2012), pp. 385–391.

[17] S. C. Chen, C. M. Liao, C. P. Chio, H. H. Chou, S. H. You, and Y. H. Cheng, *Lagged temperature effect with mosquito transmission potential explains dengue variability*

doi:10.6342/NTU202000011

in southern Taiwan: Insights from a statistical analysis, Sci. Total Environ., 408 (2010), pp. 4069–4075.

[18] L.-C. Chien and H.-L. Yu, *Impact of meteorological factors on the spatiotemporal patterns of dengue fever incidence*, Environ. Int., (2014).

[19] G. Christakos, *Modern Spatiotemporal Geostatistics*, Oxford University Press, USA, New York, 2000.

[20] G. Christakos, J. M. Angulo, and H. L. Yu, *Constructing space-time pdfs in Geosciences*, Bol. Geol. y Min., (2011).

[21] G. Christakos, R. A. Olea, M. L. Serre, H.-L. Yu, and L.-L. Wang, *Interdisciplinary Public Health Reasoning and Epidemic Modelling: The Case of Black Death*, Springer, New York, 2005.

[22] F. J. Colón-González, I. R. Lake, and G. Bentham, *Climate variability and dengue fever in warm and humid Mexico.*, Am. J. Trop. Med. Hyg., 84 (2011), pp. 757–763.

[23] N. Cressie, *The origins of kriging*, Math. Geol., (1990).

[24] E. S. Cross, L. R. Williams, D. K. Lewis, G. R. Magoon, T. B. Onasch, M. L. Kaminsky, D. R. Worsnop, and J. T. Jayne, *Use of electrochemical sensors for measurement of air pollution: Correcting interference response and validating measurements*, Atmos. Meas. Tech., 10 (2017), pp. 3575–3588.

[25] H. Q. Cuong, N. T. Vu, B. Cazelles, M. F. Boni, K. T. D. Thai, M. A. Rabaa, L. C. Quang, C. P. Simmons, T. N. Huu, and K. L. Anders, *Spatiotemporal dynamics of dengue epidemics, southern Vietnam.*, Emerg. Infect. Dis., 19 (2013), pp. 945–53.

[26] J. Dick, F. Y. Kuo, and I. H. Sloan, *High-dimensional integration: The quasi-Monte Carlo way*, 2013.

[27] J. Duncombe, A. Clements, J. Davis, W. Hu, P. Weinstein, and S. Ritchie, *Spatiotemporal patterns of Aedes aegypti populations in Cairns, Australia: assessing drivers of dengue transmission.*, Trop. Med. Int. Health, 18 (2013), pp. 839–849.

doi:10.6342/NTU202000011

[28] Edimax Inc., *AirBox : Smart Air Quality Detector with PM2.5, Temperature and Humidity Sensors*.

[29] R. M. Eggo, S. Cauchemez, and N. M. Ferguson, *Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States.*, J. R. Soc. Interface, 8 (2011), pp. 233–243.

[30] E. Esposito, S. De Vito, M. Salvato, V. Bright, R. L. Jones, and O. Popoola, *Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems*, Sensors Actuators, B Chem., 231 (2016), pp. 701–713.

[31] European Commission's Directorate General for Environment; European Environment Agency, *European Air Quality Index*.

[32] A. Fallis, *The four V's of big data*, Big Data, (2013).

[33] C. P. Farrington, H. J. Whitaker, J. Wallinga, and P. Manfredi, *Measures of disassortativeness and their application to directly transmitted infections.*, Biom. J., 51 (2009), pp. 387–407.

[34] N. M. Ferguson, M. J. Keeling, W. J. Edmunds, R. Gani, B. T. Grenfell, R. M. Anderson, and S. Leach, *Planning for smallpox outbreaks.*, Nature, 425 (2003), pp. 681–685.

[35] A. S. Fotheringham and C. Brunsdon, *Quantitative Geography : Perspectives on Spatial Data Analysis*, Sage Publications Ltd, London, 2000.

[36] G0v.tw, *g0v* 零時空汙觀測網.

[37] A. Gasparrini, *Distributed Lag Linear and Non-Linear Models in R: The Package dlnm.*, J. Stat. Softw., 43 (2011), pp. 1–20.

[38] A. Gasparrini, B. Armstrong, and M. G. Kenward, *Distributed lag non-linear models.*, Stat. Med., 29 (2010), pp. 2224–2234.

doi:10.6342/NTU202000011

[39] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, *Detecting influenza epidemics using search engine query data*, Nature, (2009).

[40] R. González, S. Infante, and A. Hernández, *Spatio-temporal hierarchical models for mapping relative risks of dengue in the Municipality of Girardot, Aragua State, Venezuela*, Boletín Malariol. y Salud Ambient., 52 (2012), pp. 33–43.

[41] Google LLC, *Google Trends*.

[42] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, *Internet of Things (IoT): A vision, architectural elements, and future directions*, Futur. Gener. Comput. Syst., 29 (2013), pp. 1645–1660.

[43] D. J. Gubler and G. G. Clark, *Dengue / Dengue Hemorrhagic Fever :*, Emerg. Infect. Dis., 1 (1995), pp. 55–57.

[44] J. D. Hamilton, *Time Series Analysis*, in Time Ser. Anal., 1994.

[45] K. H. Hampton, M. L. Serre, D. C. Gesink, C. D. Pilcher, and W. C. Miller, *Adjusting for sampling variability in sparse data: Geostatistical approaches to disease mapping*, Int. J. Health Geogr., (2011).

[46] L. Held and M. Paul, *Modeling seasonality in space-time infectious disease surveillance data.*, Biom. J., 54 (2012), pp. 824–43.

[47] Y. L. Hii, J. Rocklöv, S. Wall, L. C. Ng, C. S. Tang, and N. Ng, *Optimal Lead Time for Dengue Forecast*, PLoS Negl. Trop. Dis., 6 (2012).

[48] Y. L. Hii, H. Zhu, N. Ng, L. C. Ng, and J. Rocklöv, *Forecast of Dengue Incidence Using Temperature and Rainfall*, PLoS Negl. Trop. Dis., 6 (2012).

[49] M. Hitchman, N. Cade, T. Gibbs, and N. Hedley, *Study of the Factors Affecting Mass Transport in Electrochemical Gas Sensors*, Analyst, 122 (1997), pp. 1411–1417.

doi:10.6342/NTU202000011

[50] G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs, *A review of land-use regression models to assess spatial variation of outdoor air pollution*, 2008.

[51] D. M. Holstius, A. Pillarisetti, K. R. Smith, and E. Seto, *Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California*, Atmos. Meas. Tech., 7 (2014), pp. 1121–1131.

[52] J.-C. Huang, *Investigation of dengue fever survelliance quality in Taiwan*, tech. rep., Taipei, 2009.

[53] S. Inaida, Y. Shobugawa, S. Matsuno, R. Saito, and H. Suzuki, *The South to north variation of norovirus epidemics from 2006-07 to 2008-09 in Japan.*, PLoS One, 8 (2013), p. e71696.

[54] R. Jayaratne, X. Liu, P. Thai, M. Dunbabin, and L. Morawska, *The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog*, Atmos. Meas. Tech., (2018).

[55] W. Jiao, G. Hagler, R. Williams, R. Sharpe, R. Brown, D. Garver, R. Judge, M. Caudill, J. Rickard, M. Davis, L. Weinstock, S. Zimmer-Dauphinee, and K. Buckley, *Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States*, Atmos. Meas. Tech., 9 (2016), pp. 5281–5292.

[56] A. G. Journel, *Constrained interpolation and qualitative information-The soft kriging approach*, Math. Geol., (1986).

[57] R. Judge and R. A. Wayland, *Regulatory considerations of lower cost air pollution sensor data performance*, EM Air Waste Manag. Assoc. Mag. Environ. Manag., (2014), pp. 32–37.

[58] M. N. Karim, S. U. Munshi, N. Anwar, and M. S. Alam, *Climatic factors influencing dengue cases in Dhaka city: a model for dengue prediction.*, Indian J. Med. Res., 136 (2012), pp. 32–9.

doi:10.6342/NTU202000011

[59] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*, American Mathematical Society, the University of Michigan, 1980.

[60] D. G. Krige, *A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand*, J. Chem. Metall. Min. Soc. South Africa, (1952).

[61] K. Kuhn, D. Campbell-Lendrum, A. Haines, and J. Cox, *Using climate to predict infectious disease epidemics*, tech. rep., World Health Organization, Geneva, Switzerland, 2005.

[62] D. Lazer, R. Kennedy, G. King, and A. Vespignani, *The parable of google flu: Traps in big data analysis*, 2014.

[63] C. H. Lee, Y. B. Wang, and H. L. Yu, *An efficient spatiotemporal data calibration approach for the low-cost PM2.5 sensing network: A case study in Taiwan*, Environ. Int., (2019).

[64] W. Leiss, *Three Phases in the Evolution of Risk Communication Practice*, Ann. Am. Acad. Pol. Soc. Sci., (1996).

[65] A. R. Lifson, *Mosquitoes, models, and dengue*, 1996.

[66] Y.-C. Lin, W.-J. Chi, and Y.-Q. Lin, *The improvement of spatial-temporal resolution of PM2.5 estimation based on micro-air quality sensors by using data fusion technique*, Environ. Int., 134 (2020), p. 105305.

[67] R. Lowe, T. C. Bailey, D. B. Stephenson, R. J. Graham, C. A. S. Coelho, M. Sá Carvalho, and C. Barcellos, *Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil*, Comput. Geosci., 37 (2011), pp. 371–381.

[68] R. Lowe, T. C. Bailey, D. B. Stephenson, T. E. Jupp, R. J. Graham, C. Barcellos, and M. S. Carvalho, *The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil.*, Stat. Med., 32 (2013), pp. 864–83.

doi:10.6342/NTU202000011

[69] R. Lowe and A. M. Stewart-Ibarra, *Climate and Non-Climate Drivers of Dengue Epidemics in Southern Coastal Ecuador*, Am. J. Trop. Med. Hyg., 88 (2013), pp. 971–981.

[70] N. Masson, R. Piedrahita, and M. Hannigan, *Quantification method for electrolytic sensors in long-term monitoring of ambient air quality*, Sensors (Switzerland), 15 (2015), pp. 27283–27302.

[71] M. I. Mead, O. A. M. Popoola, G. B. Stewart, P. Landshoff, M. Calleja, M. Hayes, J. J. Baldovi, M. W. McLeod, T. F. Hodgson, J. Dicks, A. Lewis, J. Cohen, R. Baron, J. R. Saffell, and R. L. Jones, *The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks*, Atmos. Environ., 70 (2013), pp. 186–203.

[72] A. Midekisa, G. Senay, G. M. Henebry, P. Semuniguse, and M. C. Wimberly, *Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia*, 2012.

[73] A. Mohammed and D. D. Chadee, *Effects of different temperature regimens on the development of Aedes aegypti (L.) (Diptera: Culicidae) mosquitoes*, Acta Trop., 119 (2011), pp. 38–43.

[74] A. Mondini and F. Chiaravalloti-Neto, *Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city*, Sci. Total Environ., 393 (2008), pp. 241–248.

[75] A. Mukherjee, L. G. Stanton, A. R. Graham, and P. T. Roberts, *Assessing the utility of low-cost particulate matter sensors over a 12-week period in the Cuyama valley of California*, Sensors (Switzerland), 17 (2017).

[76] H. Padmanabha, F. Correa, M. Legros, H. F. Nijhout, C. Lord, and L. P. Lounibos, *An eco-physiological model of the impact of temperature on Aedes aegypti life history traits*, J. Insect Physiol., 58 (2012), pp. 1597–1608.

doi:10.6342/NTU202000011

[77] P. Reiter, *Climate change and mosquito-borne disease.*, Environ. Health Perspect., 109 Suppl (2001), pp. 141–161.

[78] S. Riley, *Large-scale spatial-transmission models of infectious disease.*, Science, 316 (2007), pp. 1298–1301.

[79] S. Riley and N. M. Ferguson, *Smallpox transmission and control: spatial dynamics in Great Britain.*, Proc. Natl. Acad. Sci. U. S. A., 103 (2006), pp. 12637–12642.

[80] S. Roberts and P. Switzer, *Mortality displacement and distributed lag models.*, Inhal. Toxicol., 16 (2004), pp. 879–888.

[81] B. K. D. Sartorius, K. Kahn, P. Vounatsou, M. A. Collinson, and S. M. Tollman, *Young and vulnerable: spatial-temporal trends and risk factors for infant mortality in rural South Africa (Agincourt), 1992-2007.*, BMC Public Health, 10 (2010), p. 645.

[82] S. Sasaki, H. Suzuki, Y. Fujino, Y. Kimura, and M. Cheelo, *Impact of drainage networks on cholera outbreaks in Lusaka, Zambia.*, Am. J. Public Health, 99 (2009), pp. 1982–1987.

[83] C.-S. Shang, C.-T. Fang, C.-M. Liu, T.-H. Wen, K.-H. Tsai, and C.-C. King, *The role of imported cases and favorable meteorological conditions in the onset of dengue epidemics.*, PLoS Negl. Trop. Dis., 4 (2010), p. e775.

[84] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, and D. Hassabis, *Mastering the game of Go without human knowledge*, Nature, (2017).

[85] M. I. T. Sloan, M. Review, M. I. T. Sloan, and M. Review, *Big Data, Analytics and the Path from Insights to Value*, MIT Sloan Manag. Rev., (2010).

[86] D. L. Smith, J. Dushoff, and F. E. McKenzie, *The risk of a mosquito-borne infection in a heterogeneous environment.*, PLoS Biol., 2 (2004), p. e368.

doi:10.6342/NTU202000011

[87] D. S. Stoffer and P. Bloomfield, *Fourier Analysis of Time Series: An Introduction*, J. Am. Stat. Assoc., (2000).

[88] Taiwan Environmental Protection Administration, *Taiwan Air Quality Monitoring Network*.

[89] H. D. Teklehaimanot, J. Schwartz, A. Teklehaimanot, and M. Lipsitch, *Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia II. Weather-based prediction systems perform comparably to early detection systems in identifying times for interventions.*, Malar. J., 3 (2004), p. 44.

[90] M. L. L. Tesar, *A Comparison of Spatial Prediction Techniques Using Both Hard and Soft Data*, PhD thesis, University of Nebraska - Lincoln, 2011.

[91] THE WORLD BANK, *Environmental Information Disclosure*, in Get. to Green - A Sourceb. Pollut. Manag. Policy Tools Growth Compet., The International Bank for Reconstruction and Development / THE WORLD BANK, Washington, DC, 2012, ch. 2.1.9, pp. 118–125.

[92] C.-T. Tsai, F.-C. Sung, P. S. Chen, and S.-C. Lin, *Exploring the spatial and temporal relationships between mosquito population dynamics and dengue outbreaks based on climatic factors*, Stoch. Environ. Res. Risk Assess., 26 (2012), pp. 671–680.

[93] J. W. Tukey, *The Future of Data Analysis*, Springer New York, New York, NY, 1992, pp. 408–452.

[94] A. M. TURING, *I.—COMPUTING MACHINERY AND INTELLIGENCE*, Mind, LIX (1950), pp. 433–460.

[95] U.S. Environmental Protection Agency, *AirNow*.

[96] US EPA, *Evaluation of Emerging Air Pollution Sensor Performance*, 2016.

[97] S. O. Vanwambeke, E. F. Lambin, M. P. Eichhorn, S. P. Flasse, R. E. Harbach, L. Oskam, P. Somboon, S. Beers, B. H. B. Benthem, C. Walton, and R. K. Butlin, *Impact of Land-use Change on Dengue and Malaria in Northern Thailand*, 2007.

doi:10.6342/NTU202000011

[98] D. C. Wheeler, *Geographically weighted regression*, in Handb. Reg. Sci., 2014.

[99] WHO, *Dengue and severe dengue*, 2013.

[100] R. Williams, V. J. Kilaru, E. G. Snyder, A. Kaufman, T. Dye, A. Rutter, A. Russell, and H. Hafner, *Air Sensor Guidebook*, Tech. Rep. 1, 2014.

[101] S. Wongkoon, M. Jaroensutasinee, and K. Jaroensutasinee, *Assessing the temporal modelling for prediction of dengue infection in northern and north-eastern, Thailand.*, Trop. Biomed., 29 (2012), pp. 339–48.

[102] P.-C. Wu, H.-R. Guo, S.-C. Lung, C.-Y. Lin, and H.-J. Su, *Weather as an effective predictor for occurrence of dengue fever in Taiwan.*, Acta Trop., 103 (2007), pp. 50–57.

[103] B. S. Xia and P. Gong, *Review of business intelligence through data analysis*, Benchmarking, (2014).

[104] H. M. Yang, M. L. G. Macoris, K. C. Galvani, M. T. M. Andrighetti, and D. M. V. Wanderley, *Assessing the effects of temperature on dengue transmission.*, Epidemiol. Infect., 137 (2009), pp. 1179–1187.

[105] ——, *Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue.*, Epidemiol. Infect., 137 (2009), pp. 1188–1202.

[106] H. Yu and D. G. Robinson, *The New Ambiguity of 'Open Government'*, SSRN Electron. J., (2012).

[107] H.-L. Yu and G. Christakos, *Spatiotemporal modelling and mapping of the bubonic plague epidemic in India.*, Int. J. Health Geogr., 5 (2006), p. 12.

[108] H. L. Yu, C. H. Lee, and L. C. Chien, *A spatiotemporal dengue fever early warning model accounting for nonlinear associations with hydrological factors: a Bayesian maximum entropy approach*, Stoch. Environ. Res. Risk Assess., (2016).

doi:10.6342/NTU202000011

[109] H.-L. Yu, S.-J. Yang, H.-J. Yen, and G. Christakos, *A spatio-temporal climate-based model of early dengue fever warning in southern Taiwan*, 2011.

[110] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, *Internet of things for smart cities*, IEEE Internet Things J., 1 (2014), pp. 22–32.

[111] A. Zanobetti, J. Schwartz, E. Samoli, A. Gryparis, G. Touloumi, J. Peacock, R. H. Anderson, A. Le Tertre, J. Bobros, M. Celko, A. Goren, B. Forsberg, P. Michelozzi, D. Rabczenko, S. P. Hoyos, H. E. Wichmann, and K. Katsouyanni, *The temporal pattern of respiratory and heart disease mortality in response to air pollution.*, Environ. Health Perspect., 111 (2003), pp. 1188–1193.

doi:10.6342/NTU202000011