A spatial analysis of provincial-level mortality for all age groups in China

Maoqi Hu

Primary Supervisor: Dr. Carmen Boado-Penas

Secondary Supervisor: Prof. Corina Constantinescu

Department of Mathematical Sciences

University of Liverpool

A thesis submitted for the degree of

Master of Philosophy

April 2022

Abstract

The life expectancy at birth in China has drastically improved over the last six decades. Compared with 43.5 years in 1960, the life expectancy at birth in 2018 was 76.7 years. However, large differences in life expectancy continue to exist within China, and age-specific mortality rates differ across provinces.

This study applies the standardised mortality ratio (SMR) to compare mortalities in mainland China for the period 2000-2015. Showing that mortality is unbalanced at the provincial level. It also explores the spatial relationship for Western and Eastern China. Global and local Moran's I indices are used to rigorously verify spatial autocorrelation and identify spatial clusters of mortality. Given the presence of positive spatial autocorrelation at the provincial level in China, a spatial panel data model is constructed with four independent variables across different dimensions: demographic, environmental, economic, and societal development. The study makes important policy recommendations in terms of improving social security and healthcare in less-developed provinces and making economic development sustainable.

Acknowledgement

I would like to thank my supervisor Dr.Carmen Boado-Penas, for providing guidance and feedback throughout this thesis. I would like to thank Prof. Corina Constantinescu, for her tremendous encouragement throughout my study and life. Special thanks to Prof. Ana Debon and Dr. Patricia Carracedo for the considerate guidance my study.
Many thanks to all staffs and students at the Institute for Financial and Actuarial Mathematics (IFAM). Each one of them has been there to support me whenever I need any help. I would like to thank Prof. Jiro Akahori for inviting me to visit the Ritsumeikan University. I would like to AIMS Rwanda for giving me a lot of inspiration in my academic and daily life.
I would also like to thank Dr. Sule Sahin for agreeing to be my internal examiner. Thanks for my external examiner, Prof. Séverine Arnold, for attending my first presentation and will attend my last presentation.
Many many thanks for my parents for all their support, no matter what I decide, they always stay by my side. I would also like to thank my husband Wei Zhu, for supporting me from all aspects in life.

Contents

Chapter 1

Introduction

Studies on mortality behaviour in China usually focus on the elderly. This study focuses on mortality rates at the provincial level for all age groups and populations between 2000 and 2015 to analyse China's population distribution and highlight provinces with high mortality. The study period reflects a dramatic boost to China's economy. For instance, GDP per capita increased eight-fold from 2000 to 2015, a change that had occurred over three decades (1970–2000) in the past. Parallelly, economic development across China was inconsistent, which enlarged the gap between provinces. Studying the spatial mortality differences at the provincial level along with its impact factor variables helps identify the geographic characteristics of mortality in China.

A decline in mortality rates, coupled with an improvement in life expectancy, is increasingly being observed worldwide. However, the differences in life expectancy between various countries and areas remain significant. Particularly, life expectancy at birth exceeded 80 years in Europe, North America, Australia, and New Zealand as of 2019. However, the life expectancy at birth in other parts of the world, such as sub-Saharan Africa, was only around 61 years. The life expectancy at birth was 76.5 years in East and South-East Asia and 69.9 years in Central and South Asia (Nations, 2019). Furthermore, although the world sex ratio was close to 1, females outnumbered males in the older age groups, because of longer average life expectancy. The life expectancy at birth for the total global population was 72.6 years in 2019, wherein the male and female life expectancies at birth were 70.2 and 75.0 years, respectively. (Nations, 2019). The life expectancy at birth in China is also unbalanced between sexes. As an important index of population health, China's life expectancy at birth increased from 43.72 years in 1960 to 75.92 years in 2015[1]. Specifically, the life expectancies of males and females in 2015 were 73.79 and 78.29 years, respectively. However, there are differences in life expectancy at the provincial level. For example, in 2015, the life expectancy at birth in Beijing was 83.5 years, whereas that in Qinghai was 69.4 years (Zhou et al., 2016). Several factors affect mortality rates and life expectancy directly and indirectly at the provincial level in China. This thesis examine these factors from the demographic, environmental, and economic points of view:

1) Demographically, the age distribution of population has changed in China, both at the national and provincial levels. Since the implementation of the Central Government's one-child policy[2] in the late 1970s, the fertility rate has dropped sharply, and population growth has been controlled (Coale, 1981; Bongaarts and Greenhalgh, 1985). Subsequently, the dominant age structure of the population has shifted from working adults to the elderly. China was recognised as an ageing society in 2000 when the proportion of the population above 65 years of age exceeded 7% for the first time. At the provincial level, domestic migration from less-developed provinces to more-developed provinces changed the demographic distribution in different provinces (Shen,

[1] Data source: World Bank

[2] It was announced in late 2015 that the program would end in early 2016.

2013). Over the last four decades, many people in China, including millions of parents from small cities or rural areas, have migrated to larger cities for jobs, leaving their children (termed ‘left-behind children’) with grandparents. This phenomenon further diversifies the age-specific population distribution between provinces as well as urban and rural areas.

2) In terms of environmental aspects, industrial activity leads to pollution, while green spaces reduce pollution. Cao et al. (2012) stated that the growth of the heavy and construction industries in China brought about a substantial increase in energy consumption and polluting emissions. Construction investment increased dramatically during China’s economic boom decades, which led to a structural transformation from agriculture and consumption manufacturing to construction, heavy industry, and export-related manufacturing. Inevitably, the transformation generated substantial air pollutants, which damaged urban residents’ health. As industrialisation in China is unbalanced, different provinces are differently affected by air pollution (Xu and Lin, 2016, 2018; Xie et al., 2019; Zhang et al., 2020). Particulate Matter (PM) 10 (PM_{10}), nitrogen dioxide (NO_2) and sulphur dioxide (SO_2) are the main air pollutants routinely measured by the Chinese government. Specifically, China is the third-largest SO_2 emitting country in the world, because the largest SO_2 source is coal consumption, which continues to dominate China’s energy consumption, accounting for 57.7% of the total energy consumption in 2019[3]. This has put the spotlight on the Yangtze River and Beijing–Tianjin areas, as the provinces in these areas have comparatively higher levels of SO_2, which is associated with increased premature mortality and morbidity (Cao et al., 2012). In addition, SO_2 has several negative effects on human health, leading to respiratory issues, pulmonary oedema, eye

[3] Data source: The National Bureau of Statistics (NBS)

irritation, asthma, and cardiopulmonary diseases (Lin et al., 2004; Khan and Siddiqui, 2014; Goudarzi et al., 2016). To improve the residential environment and reduce air pollution, green spaces have been increased in several provinces to benefit human health (Twohig-Bennett and Jones, 2018), especially in the urban areas (Alcock et al., 2014; Gascon et al., 2015; Kondo et al., 2018), thus affecting mortality rates. Wang and Tassinary (2019) proposed that the spatial distribution of green spaces in cities is obviously correlated with the mortality risk. Zhang et al. (2020) found that although China has been steadily working on urban greening, large inter-provincial differences still exist.

3) Economic development and urbanisation levels differ from province to province. Démurger et al. (2002) pointed out strong economic inequalities in countries with vast geographies. China is the third-largest country by land area in the world, and its landscapes vary significantly across its expansive territory. Economic inequalities have been observed between China's inland and coastal provinces (Hao and Wei, 2010; Chen et al., 2017). Like many other countries, China initiated economic reforms in the coastal provinces, giving them an early advantage in terms of economic development. Cutler et al. (2006) found that it is easier to provide public health infrastructure in areas with high income rather than in poorer areas. Zhao (2006) stated that mortality rates in China are considerable unbalanced between the more-developed provinces and the less-developed provinces. Additionally, the former are usually more urbanised because of constructive government policies that were implemented in the mid-1980s (Wan, 2008). Hence, urbanisation varies at the provincial level in China (Lin et al., 2018) and influences regional mortality rates (Luo et al., 2015; Li et al., 2016; Hou et al., 2019).

Existing data analysis

Owing to limited data, most research on China's mortality rates has been conducted at the national level (Banister and Hill, 2004; Zhao, 2012; Zhao et al., 2013; Huang and Browne, 2017; Li et al., 2019); fewer studies focus on mortality rates at the provincial level in China. Ren et al. (2004) analysed provincial changes in mortality rates by building new life tables for different provinces. Lu et al. (2019) designed a Bayesian hierarchical framework based on principal components and a random walk process to estimate and forecast mortality rates in China at the provincial level.

Some studies analyse the spatial autocorrelation of mortality in China. Wang et al. (2015) analysed spatial autocorrelation of the human lifespan in China for the years 1990, 2000 and 2010. Chen et al. (2019) showed the differences between the regional ageing populations in China from temporal and spatial perspectives over the period 1998–2014. Yang et al. (2020) found that the unbalanced spatial distribution of socio-economic development leads to unequal health situations among the elderly at the national level. Wu et al. (2020) analysed the spatial clustering characteristics of life expectancy in China for the year 2010.

A spatial dependence model examines the relationship between mortality and impact factors in China by explaining the behaviour of inter-related geographical units. Xiang and Song (2016) focused on the spatial analysis of perinatal mortality at the provincial level in China between 1996 and 2013. Wu et al. (2019) found spatial differences in China's ageing population for the period 2000–2010 driven by demographic factors. However, the spatial dependence model does not control for spatial and temporal heterogeneity (Arellano, 2003). The spatial panel data model is applied to control for spatial and temporal heterogeneity. Li (2017) used the spatial panel data model to conclude that China's ageing population at the provincial level is impacted by the gross regional product (GRP), medical treatment, etc. Luo et al. (2018) analysed sanitation at the provincial level in China between 2006 and 2015 using a spatial panel data model, because sanitation plays an importation role in disease prevention.

This study identifies which factors affect mortality rates in China by looking into differences in mortalities between different provinces. It is the first research project that applies a spatial panel data model on mortality rates in China for all age groups in different years. This study helps the government narrow the provincial disparity in mortality rates and pay more attention to those provinces that have high mortalities. In addition, the study interprets the causes of death from four aspects and provides references for the improvement of life expectancy in each province.

This study used data on all age groups to analyse mortality rates at the provincial level for China between 2000 and 2015. The remainder of this thesis is organised as follows. Chapter 2 presents a literature review on mortality standardisation and the spatial panel data model. Chapter 3 introduces mortality standardisation and spatial autocorrelation methods. Chapter 4 presents the spatial panel data model, which is used to examine the factors that impact mortality in China. Chapter 5 presents the main results of the mortality cluster and spatial panel data methods. Chapter 6 presents the conclusions. The appendix outlines all the data sources and the ones that will be used in the thesis. Additionally, the mortality data, accounting for the missing values, and the impact factor variables that explain mortality behaviour are also described.

Chapter 2

Literature review

This study is based on mortality standardisation methods, spatial econometrics, and spatial panel data models. Mortality standardisation allows the comparison of mortality across various age distributions, a technique that has been widely applied in the past in many fields. The standardised mortality ratio (SMR) and comparative mortality figure (CMF) are generally used for mortality analysis and comparisons. In this study, the spatial panel data model was used to examine the factors that impact mortality.

2.1 Mortality standardization

Most studies on standardisation have emphasised its applications in medical science (Benjamin, 1968; Miettinen, 1972; Lilienfeld, 1978; Logan, 1982). As many chemicals used in modern society have potential health hazards, Jarup (2004) assessed the risk of chemically induced diseases. Woolf et al. (2004) analysed the health impact of resolving racial disparities based on U.S. mortality data. Gächter and Theurl (2011) examined health status at the local level in Austria for the period 1969–2004.

The standardised calculation of death numbers in actuarial mathematics dates to the 18th century (Dale, 1777; Tetens, 1786). Crude death rate comparisons were not always accurate because they only described the proportion of deaths in the population for a specific period in each geographical area, without considering the age distribution for such numbers. William Farr, who had been working at the Office of the Registrar General for England and Wales since 1839, introduced age-specific death rates in 1841 to analyse changes in mortality rates for different age groups in England and Wales (General, 1841a). Unfortunately, this method was complicated when applied to a large number of age groups for deaths and the population. Hence, several contemporaneous methods were developed to measure mortality rates and make comparisons between the studied and standard populations. Such techniques were called standardisations. The concept of a standard population was first introduced in the Registrar General's report (General, 1853) and defined in terms of a set of 'healthy' countries with crude death rates less than 1.7%. William Farr took the age structure of death rates into account and applied the standard population's age-specific death rates to other countries' populations. This was the first form of standardised death rates (General, 1857). Indirect standardisation, a unique age-specific method, was the most widely-applied index until the SMR was introduced in 1883. As age-specific death numbers were not always available for the study population, the Office of the Registrar General for England and Wales proposed direct standardisation in 1883 (General, 1883). Subsequently, the CMF, an index of direct standardisation, was proposed in 1884 (General, 1884). Indirect and direct standardisations are the most widely-applied methods for analysing and comparing death numbers/rates. Other standardised measures include standardised rates, such as the equivalent average death rate (Yule, 1934), cumulative rate and comparative mortality rate, and standardised ratios, such as the Yule's index (Yule, 1934), comparative mortality index(General, 1841b) and Fisher's Ideal Index (Fisher, 1927).

Yule (1934) was the first to derive the standard error of standardised rates. Sampling errors of mortality statistics were thoroughly discussed by Westergaard (1882). Rubin and Westergaard (1886) introduced occupational mortality, which is an early application of standardisation as an analytical statistical methodology. Breslow and Day (1975), Breslow (1975), Breslow and Day (1985) and Hoem (1987) considered the SMR as a maximum likelihood (ML) estimator in a proportional hazards model.

As statistical methods, indirect and direct standardisation (Fleiss et al., 2013; Inskip, 2014; Inskip et al., 1983) are widely used in many fields, including medical studies. Doll and Cook (1967) illustrated how the standard population influences age-standardised incidence rates of cancer, and Day (1976) used direct standardisation to analyse cancer incidence rates. The relationship between moderate arsenic levels and 23 specific diseases was analysed for South-eastern Michigan by using the SMR (Meliker et al., 2007). Tseng et al. (2011) used the SMR in a study of inpatient suicides at a general hospital; Mok et al. (2013) studied the effect of renal disease on the SMR and life expectancy of patients with systemic lupus erythematosus. Kashyap et al. (2019) described the association between septic shock definitions and standardised mortality ratios for a contemporary cohort of critically ill patients. There are many related studies for China, mostly in the epidemiological and medical fields (Lai et al., 2000; Ding et al., 2006; Mok et al., 2011; Zhu et al., 2012; Du et al., 2012; Zhang et al., 2018; Liu et al., 2018). Apart from their applications in the medical field, Tripepi et al. (2010) used both direct and indirect standardizations to compare drinkers' death rates. Sugawara et al.

(2013) found the relationship between lithium in tap water and suicide mortality in Japan using the SMR. Carracedo and Debón (2016) applied the SMR to compare mortality levels in Europe from 1990 to 2009. Kim et al. (2020) analysed mortality rates in Korea using the SMR, CMF, and life expectancy.

2.2 The spatial panel data model

This study also applies the spatial panel data model in spatial econometrics. Spatial econometrics analyses spatial interaction effects between geographical locations, such as countries, provinces, or regions, and is usually applied to explain economic behaviour (Elhorst, 2014). In the past, researchers interested in spatial interactions between geographical locations considered a spatial weights matrix W to describe such spatial arrangements. Haining (1993); Anselin and Bera (1998); Arbia (2006); LeSage and Pace (2009); Cressie (2015) made key contributions in this field. Spatial dependence models can be estimated by ML estimation (Ord, 1975), Bayesian methods (LeSage, 1997), generalised method of moments (GMM) (Kelejian and Prucha, 1998), and quasi-maximum likelihood (QML) (Lee, 2004).

The most popular spatial dependence models are the spatial lag and spatial error models. Both include only one type of interaction effect[1]. Models with more than one interaction effect have attracted more interest since 2007. Harry Kelejian (Kelejian and Prucha, 1998) introduced a model (renamed as the Kelejian–Prucha model by Elhorst (2010)) with both endogenous interaction effects and interaction effects among the error terms at the first World Conference of the Spatial Econometrics Association. In the 54th North American Meeting of the Regional Science Association International in 2010, James Le Sage advocated for a model

[1] For details of the interaction effects, see Section 4.1

with endogenous interaction effects.

Over the last two decades, more studies have focused on spatial panel data models, which extend the cross-sectional spatial dependence model by using panel data. Panel data is the combination of time series and cross-sectional data. Baltagi et al. (2003) were the first to consider the testing of spatial interaction effects in a spatial panel data model. Anselin (2013) also discussed the spatial dependence panel model in detail. The spatial dependence panel model has advantages such as more degrees of freedom and increased efficiency of estimation. More importantly, it performs well for more complicated cases. Lagrange multiplier (LM) tests (Burridge, 1980; Anselin, 2013) and robust LM tests (Anselin et al., 1996) were introduced to estimate spatial panel data models. Baltagi et al. (2003) considered several LM tests in panel data models with spatial error correlation. Additionally, Anselin and Hudak (1992) focused on ML estimation of the spatial lag model. Driscoll and Kraay (1998) and Bell and Bockstael (2000) estimated the spatial panel data set using the GMM. Lee and Yu (2012) introduced QML estimation of spatial dynamic panel data models with a time-varying spatial weights matrix.

There are several studies on the applications of spatial panel data models. Using these models, Gwatkins et al. (2007), Preston (1975) and Preston (1980) found that people with higher incomes have better health status and lower mortality. Elhorst (2003) estimated a spatial panel data model for both fixed and random effects. Druska and Horrace (2004) and Baylis et al. (2011) applied spatial approaches to panel data in agricultural economics. Mutl and Pfaffermayr (2011) considered a Cliff–Ord-type spatial lag model and estimated instrumental variables with fixed and random effects. Pfaffermayr (2009) and Wang and Lee (2013) estimated spatial panel data models with missing data. Barufi et al. (2012) applied the spatial panel data model to analyse infant mortality in Brazil. A spatial Durbin model was used to study U.S. mortality rates (Yang et al., 2015). Similar analysis has been performed for Europe by Carracedo and Debón (2016). Wang and Luo (2018) showed how heating energy utilisation impacts life expectancy in China. Furthermore, the spatial panel data model can also be used in transport research (Frazier and Kockelman, 2005), social economics (Egger et al., 2005) and products analysis (Baltagi and Li, 2006).

Chapter 3

The spatial autocorrelation of mortality

Standardised methods of mortality include the SMR, CMF, indirectly standardised rate (ISR), and directly standardised rate (DSR). This chapter provides a comprehensive comparison between the most two commonly-used models, SMR and CMF. The CMF was found to be more sensitive to numerical instabilities than the SMR for one or two of the age-specific rates. After quantifying mortality, the global and the local Moran indices were used to assess the spatial autocorrelation of mortality between provinces. Furthermore, the local Moran index was also applied to analyse mortality spatial clusters and spatial outliers.

3.1 Mortality standardisation

It is difficult to compare provinces with dissimilar age distributions of populations and deaths. Mortality standardisation helps resolve this problem. Quantifying mortality also enables spatial autocorrelation analysis. This section introduces the SMR, CMF, ISR, and DSR, which are widely used in practice owing to their good performance in comparing groups with varying population sizes. The SMR and CMF are also compared. Table 3.1 provides a list of notations that have been subsequently used in the equations for standardised rates and ratios.

Description	Study population (province i year t)	Standard population (year t)
Population in age group g	$n_{i,g}(t)$	$N_g(t)$
Total population	$n_i(t) = \sum_g n_{i,g}(t)$	$N(t) = \sum_g N_g(t)$
Deaths in age group g	$d_{i,g}(t)$	$D_g(t)$
Total deaths	$d_i(t) = \sum_g d_{i,g}(t)$	$D(t) = \sum_g D_g(t)$
Crude death rate	$r_i(t) = \frac{d_i(t)}{n_i(t)}$	$R(t) = \frac{D(t)}{N(t)}$
Age-specific death rate in age group g	$r_{i,g}(t) = \frac{d_{i,g}(t)}{n_{i,g}(t)}$	$R_g(t) = \frac{D_g(t)}{N_g(t)}$

Table 3.1: Notations for standardised mortality measures

3.1.1 Indirect and direct mortality standardisation

The standardized mortality ratio (SMR) is expressed as the ratio of the number of actual deaths in the study population to the number of expected deaths:

$$\text{SMR} = \frac{\text{Actual deaths (in study population)}}{\text{Expected deaths (in study population)}}. \tag{3.1}$$

The expected number of deaths is measured by the standard age-specific death rate times the study population. The standard population implies China's national population in the present study. Hence, the standard age-specific death rate is defined as the ratio of the number of national age-specific deaths to the national age-specific population (see Table 3.1). In addition, the expected deaths in each age group depend on the age-specific study population $n_{i,g}(t)$. For province i in year t, the SMR is represented as:

$$\mathrm{SMR}_{it} = \frac{\sum_{g=1}^{k} n_{i,g}(t)\frac{d_{i,g}(t)}{n_{i,g}(t)}}{\sum_{g=1}^{k} n_{i,g}(t)\frac{D_g(t)}{N_g(t)}} = \frac{\sum_{g=1}^{k} d_{i,g}(t)}{\sum_{g=1}^{k} n_{i,g}(t)\frac{D_g(t)}{N_g(t)}} = \frac{d_i(t)}{\sum_{g=1}^{k} n_{i,g}(t)R_g(t)}. \tag{3.2}$$

The standard error (SE) of the SMR is (Breslow et al., 1980; Armitage et al., 2008),

$$\mathrm{SE}(\mathrm{SMR}_{it}) = \frac{\sqrt{d_i(t)}}{\sum_{g=1}^{k} n_{i,g}(t)\frac{D_g(t)}{N_g(t)}}. \tag{3.3}$$

Logarithmic transformation when constructing test statistics or confidence intervals helps reduce the skewness in the SMR and improves the approximation to normality of the test statistic distribution. Hence,

$$\mathrm{SE}(\mathrm{logSMR}_{it}) = \frac{\mathrm{SE}(\mathrm{SMR}_{it})}{\mathrm{SMR}_{it}} = \frac{1}{\sqrt{d_i(t)}}. \tag{3.4}$$

The comparative mortality figure (CMF) is defined as the ratio of the expected number of deaths in the standard population to those observed.

$$\mathrm{CMF} = \frac{\text{Expected deaths (in standard population)}}{\text{Actual deaths (in standard population)}}. \tag{3.5}$$

The CMF is also called the standardised rate ratio (SRR), which is as follows:

$$\mathrm{CMF}_{it} = \mathrm{SRR}_{it} = \frac{\sum_{g=1}^{k} N_g(t)\frac{d_{i,g}(t)}{n_{i,g}(t)}}{\sum_{g=1}^{k} N_g(t)\frac{D_g(t)}{N_g(t)}} = \frac{\sum_{g=1}^{k} N_g(t)\frac{d_{i,g}(t)}{n_{i,g}(t)}}{D(t)}. \tag{3.6}$$

The standard error (SE) of the CMF is (Breslow et al., 1980; Armitage et al., 2008),

$$\mathrm{SE}(\mathrm{CMF}_{it}) = \frac{\sqrt{\sum_{g=1}^{k} N_g^2(t)\frac{d_{i,g}(t)}{n_{i,g}^2(t)}}}{D(t)}. \tag{3.7}$$

Similarly, the CMF should be logarithmically transformed:

$$\mathrm{SE}(\log\mathrm{CMF}_{it}) = \frac{\mathrm{SE}(\mathrm{CMF}_{it})}{\mathrm{CMF}_{it}} = \frac{\sqrt{\sum\limits_{g=1}^{k} N_g^2(t)\frac{d_{i,g}(t)}{n_{i,g}^2(t)}}}{\sum\limits_{g=1}^{k} N_g(t)\frac{d_{i,g}(t)}{n_{i,g}(t)}}. \tag{3.8}$$

If the value of the SMR (or the CMF) is greater than one, it represents a disadvantageous mortality experience, as the mortality rate of the study population exceeds that of the standard population, that is, more deaths are observed than expected. Conversely, an advantageous mortality experience is represented by an SMR (or CMF) value of less than one.

Besides the SMR and CMF, there are two kinds of standardised rates that can be expressed in terms of the SMR and CMF. The ISR is the expected mortality rate in the study population under the assumption of age-specific mortality rates in a standard population. It is calculated as

$$\mathrm{ISR}_{it} = \frac{d_i(t)R(t)}{\sum\limits_{g=1}^{k} n_{i,g}(t)R_g(t)} = \mathrm{SMR}_{it} \cdot R(t). \tag{3.9}$$

The directly standardized rate (DSR) is defined in a symmetric way, wherein the age-specific death rates in the study population are applied to the standard population.

$$\mathrm{DSR}_{it} = \sum_{g=1}^{k} \frac{N_g(t)}{N(t)} \frac{d_{i,g}(t)}{n_{i,g}(t)} = \mathrm{CMF}_{it} \cdot R(t). \tag{3.10}$$

Yule (1934) introduced the weight concept in the mortality standardisation system, wherein a weight $w_{i,g}(t)$ is allocated to age group g in year t in province i through various standardisation methods. Using different forms of weights, the SMR and CMF can be expressed as $\frac{\sum w_{i,g}(t) r_{i,g}(t)/R_g(t)}{\sum w_{i,g}(t)}$ (Inskip et al., 1983). The ISR and

DSR can be rewritten as $\sum w_{i,g}(t)r_{i,g}(t)$. The different forms of $w_{i,g}(t)$ and the corresponding death rates are outlined in Table 3.2.

Ratio and rates	$w_{i,g}(t)$	Formula
Indirectly standardized death rate (ISR$_i(t)$)	$\frac{R(t)n_{i,g}(t)}{\sum R_g(t)n_{i,g}(t)}$	$\frac{\sum R(t)n_{i,g}(t)r_{i,g}(t)}{\sum n_{i,g}(t)R_g(t)}$
Directly standardized death rate (DSR$_i(t)$)	$\frac{N_g(t)}{N(t)}$	$\sum \frac{N_g(t)}{N(t)} \frac{d_{i,g}(t)}{n_{i,g}(t)}$
Standardized mortality ratio (SMR$_i(t)$)	$R_g(t)n_{i,g}(t)$	$\frac{\sum r_{i,g}(t)n_{i,g}(t)}{\sum R_g(t)n_{i,g}(t)}$
Comparative mortality figure (CMF$_i(t)$)	$D_g(t)$	$\sum \frac{D_g(t)r_{i,g}(t)}{D(t)R_g(t)}$

Table 3.2: Different standardised ratios and rates in terms of weight

3.1.2 Comparison of the SMR and CMF

Both the SMR and CMF are popular measures and have pros and cons that make them suitable for specific cases. The CMF ensures consistency between all study populations, because every study population is standardised using the same population. There is no need for every age-specific death rate to be larger in region A compared to region B. If they are equivalent for all age groups but one (where the value in that age group is higher in A than in B for example), the value of CMF in A will be higher than in B. In contrast, the SMR does not involve a direct comparison because it is not based on the same standard population. The SMR comparison is based on expected deaths, which considers age-specific death rates in standard population $R_g(t)$ (Julious et al., 2001). The CMF is only applicable when all age-specific death numbers are known for the study population. If any value in the study population $d_{i,g}(t)$ is unavailable, it leads to a large error in estimation (Inskip, 2014). However, the SMR has an advantage in this situation as its calculation only requires the total number of deaths in the study population $d_i(t)$, rather than the age-specific deaths $d_{i,g}(t)$.

Additionally, the CMF is unstable when the age-specific death rates $\frac{d_{i,g}(t)}{n_{i,g}(t)}$ are based on small age-specific number of deaths $d_{i,g}(t)$. Assume that there are three age groups whose standard populations are 500, 6000, and 300, and the total standard number of deaths is 300. The study populations are 100, 600, and 1, and the number of deaths are 2, 5, and 1, respectively, in the study population. According to Equation (3.6),

$$\text{CMF} = \frac{500 \times (2/100) + 6000 \times (5/600) + 300 \times (1/1)}{300} = 1.2$$

However, if the individual in the last age group is alive, the same calculation yields a large difference compared with 1.2.

$$\text{CMF} = \frac{500 \times (2/100) + 6000 \times (5/600) + 300 \times (0/1)}{300} = 0.2$$

In addition, the CMF values rely to a large extent, on the age structure of the standard population. When the older age groups in the standard population $N_g(t)$ have heavy weights, the CMF value is large because the age-specific rate $\frac{d_{i,g}(t)}{n_{i,g}(t)}$ is higher than the others. For example, their population is changed such that the standard age-specific populations are 500, 6000 and 400, we obtain

$$\text{CMF} = \frac{500 \times (2/100) + 6000 \times (5/600) + 400 \times (1/1)}{300} = 1.53$$

Thus, practically, the CMF is more sensitive than the SMR to numerical instabilities in one or two of the age-specific rates. The SMR can be considered as a weighted average of the ratios of age-specific mortality rates for the study and standard populations, wherein the weights (see Table 3.2) minimise the variance of the weighted average (Breslow, 1987). The SE also demonstrates the numerical stability of the SMR. Equations (3.4) and (3.8) also show that the SE of the SMR only depends on the changes in the number of total death $d_i(t)$, but that of the CMF depends on the age-specific number of deaths $d_{i,g}(t)$. In general, the SMR has a smaller SE than the CMF. Moreover, if the study population and standard population distributions are different, the SMR and CMF differ, as the SMR is the ML estimate (Armitage et al., 2008).

3.2 Spatial autocorrelation

Once the appropriate method of mortality standardisation is chosen, global and local Moran indices are used to assess the spatial autocorrelation of mortality between a province and its neighbourhoods (Anselin, 1995). The existence of spatial autocorrelation establishes that a spatial panel data model should be used to analyse mortality.

The global Moran's I is applied to confirm the presence of spatial autocorrelation and measure the global autocorrelation across all provinces (Moran, 1950a,b). The null hypothesis in this case is that there is no spatial autocorrelation. The local Moran's I is used to assess the influence of individual provinces and identify cluster and outlier values. Similar to that in the global Moran's I, the null hypothesis in the local Moran's I is that there is no local spatial association (Cliff and Ord, 1981).

There are three types of possible results for the global Moran's I—spatially positive correlation, spatially negative correlation, and spatial independence. A positive global Moran's I value shows positive autocorrelation, whereas a negative value indicates negative autocorrelation. Positive autocorrelation implies that when the standardised mortality value (SMR or CMF) of a province increases or decreases, the SMR values for its neighbours also increase or decrease, respectively. In the negative case, a province's standardised mortality value (SMR or CMF) is inversely related to its neighbours' values. When the value of Moran's I equals zero, there is no spatial autocorrelation. Global Moran's I is defined as:

$$\begin{aligned} GM_t &= \frac{I \sum_i \sum_j w_{ij}(y_{it} - \bar{y}_t)(y_{jt} - \bar{y}_t)}{\sum_i \sum_j w_{ij}(y_{it} - \bar{y}_t)^2}, \\ \bar{y}_t &= \frac{1}{I} \sum_{i=1}^{I} y_{it}, \\ &\text{for } i \in \{1, ..., I\}, \quad j \in \{1, ..., I\} \text{ and } i \neq j \end{aligned} \tag{3.11}$$

where I is the total number of provinces in mainland China, y_{it} is the value of standardised mortality (the SMR or CMF in the study) of province i in year t. $\bar{y}_t$ is the average of y_{it} in all provinces in year t, and w_{ij} is the ijth element of the spatial weight matrix W.

The spatial weight matrix W used to measure spatial relations between different provinces, is an $I \times I$ matrix with positive and deterministic elements w_{ij}. The values of w_{ij} aare usually based on distance functions (such as Euclidean metrics) or the spatial distance/neighbour. The spatial weight matrix in the study is chosen as a first-order binary contiguity matrix (Anselin, 1995), which means that provinces are assumed be influenced only by their neighbours. Hence, the elements of the spatial weight matrix W take the following values:

$$w_{ij} = \begin{cases} 0, & \text{if } j \notin L(i), \\ \frac{1}{L}, & \text{if } j \in L(i), \\ 0, & i = j = 1, ..., I \end{cases} \tag{3.12}$$

where L is the total number of neighbours of province I and $L(i)$ is the set of neighbours of province i. The spatial weight matrix is row-standardised, which means that each row totals to unity.

The local Moran's I measures the local spatial correlation between a given province and its surrounding provinces (Anselin, 1995). Apart from the local Moran's I, the Moran scatter plot is also used to measure the local spatial autocorrelation between a given province and its neighbours (Anselin, 1996). The local Moran index LM_{it} is defined as

$$\begin{aligned} LM_{it} &= \frac{(y_{it} - \bar{y}_t)}{S^2(y_t)} \sum_j w_{ij}(y_{jt} - \bar{y}_t), \\ S^2(y_t) &= \frac{1}{I} \sum_j (y_{jt} - \bar{y}_t)^2, \\ &\text{for } i \in \{1, ..., I\}, \quad j \in \{1, ..., I\} \text{ and } i \neq j, \end{aligned} \tag{3.13}$$

where $S^2(y_t)$ is the variance of y_t at time t.

The local Moran's I provides a statistic for each province with an assessment of significance. Meanwhile, it shows that the sum of the local Moran's I is proportional to the global Moran's I (Anselin, 1995).

Different values of the local Moran's I divide all provinces into four classes: HH, HL, LH, and LL, as summarised in Table 3.3.

Class	Standardised mortality for province i	Standardised mortality for the neighbour of province i
HH	Above mean	Above mean
HL	Above mean	Below mean
LH	Below mean	Above mean
LL	Below mean	Below mean

Table 3.3: Classification of provinces based on the local Moran's I

Classes HH and HL represent provinces with high values of standardised mortality (the SMR or CMF) surrounded by neighbours with high or low values, respectively. Conversely, classes LH and LL represent provinces with low values (of the SMR or CMF) surrounded by neighbours with high or low values, respectively.

For significant values of standardised mortality, it can be concluded that a spatial cluster and spatial outliers exist. The spatial cluster has a positive local Moran's I, which means its standardised mortality and that of its neighbours are either both above or both below the mean (classes HH or LL). Conversely, the spatial outlier has a negative local Moran's I, and the province belongs to the classes HL or LH.

Chapter 4

The spatial panel data model

When spatial autocorrelation of mortality exists at the provincial level in China, the spatial panel data model allows the analysis of mortality differences from various perspectives, such as demographic, environmental, and economic. The spatial panel data model can involve spatial effects, as well as provide information on the spatial relationship between variables, allowing further understanding of mortality differences within China. This chapter gives an overview of the linear spatial dependence models and their interrelationships. When the cross-sectional spatial dependence model is extended to panel data, it is called a spatial panel data model. This chapter also outlines the two types of spatial panel data models.

4.1 The spatial dependence model

The spatial dependence model exhibits three types of interaction effects—endogenous interaction effects within the dependent variable, interaction effects among the error terms, and exogenous interaction effects among the independent variables. The spatial lag model, also called the spatial autoregressive (SAR), contains the endogenous interaction effects within the dependent variable. The spatial error model (SEM) includes the interaction effects among the error terms. In addition, the exogenous interaction effects among the independent variables are shown in the spatially lagged X model (SLX). In practice, the SAR and SEM are used in spatial dependence analysis.

To understand the spatial dependence model, we begin with the non-spatial linear regression model,

$$y_i = \alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i \quad \text{where } i = 1, ..., I, \tag{4.1}$$

where y_i is the dependent variable for province i, and α is a constant parameter to be estimated. $\boldsymbol{x}_i$ is $K \times 1$ vector of exogenous explanatory variables for the ith province, $\boldsymbol{\beta}$ is a $K \times 1$ coefficient vector to be estimated, the disturbance term ε_i is independently and identically distributed for all i with zero mean and σ^2 variance[1].

An endogenous interaction effect implies that a change in the dependent variable of one province relies on the change in the dependent variables of the other provinces. In other words, the value of the province's dependent variable is jointly determined with those of the neighbouring provinces' dependent variables (Elhorst, 2014). The spatial lag model capturing the endogenous interaction effects thus takes the form

$$y_i = \lambda \sum_{j}^{I} w_{ij} y_j + \alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i \tag{4.2}$$

where $\sum_{j}^{I} w_{ij} y_j$ is the endogenous interaction effect among dependent variables, and λ is the spatial autoregressive coefficient.

The exogenous interaction effect indicates that the change in the dependent variable of one province relies on the change in the independent explanatory variables

[1] Another specification for the disturbances is considered in Kapoor et al. (2007).

of the other provinces. Similarly, the SLX model takes the form

$$y_i = \theta \sum_j^I w_{ij} \boldsymbol{x}_j + \alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i \tag{4.3}$$

where $\sum_j^I w_{ij} \boldsymbol{x}_j$ represents the exogenous interaction effects among independent variables, and θ represents a fixed but unknown parameter to be estimated.

In the SEM, the error term for one province is assumed to depend on the error terms for other provinces:

$$\begin{aligned} y_i &= \alpha + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \xi_i \\ \xi_i &= \rho \sum_j^I w_{ij} \xi_j + \varepsilon_i, \end{aligned} \tag{4.4}$$

where $\sum_j^I w_{ij} \xi_j$ represents the interaction effects among the disturbance terms and ρ is the spatial autocorrelation coefficient.

Each of the above models contains only one type of interaction effect. Some other models have more than one interaction effect, such as the spatial Durbin model, which includes endogenous interaction effects among the dependent variables and exogenous interaction effects among the independent variables. Figure 4.1 summarises the relationships between all the spatial dependence models (a similar figure in the matrix form is found in Elhorst (2014))

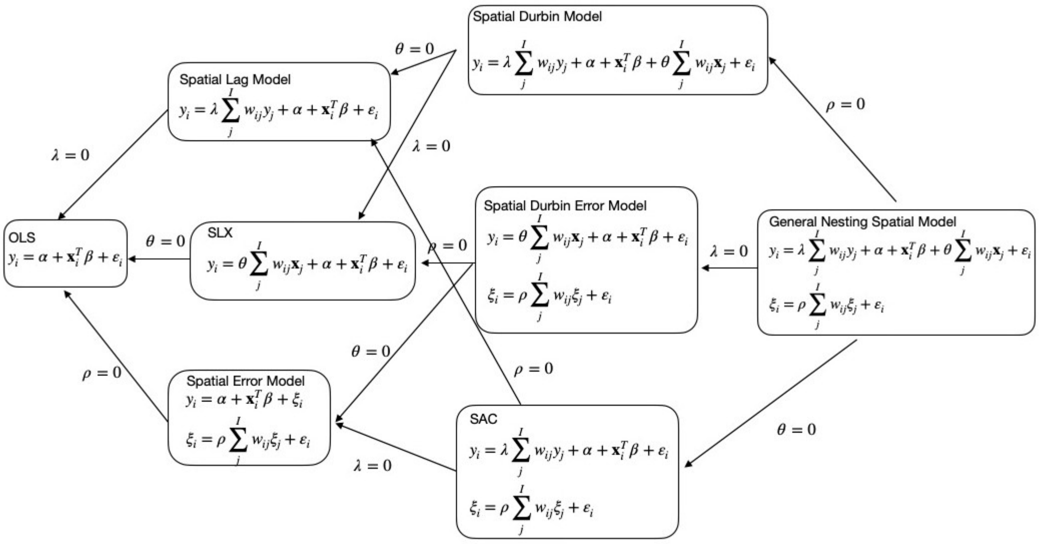

Spatial Durbin Model
$y_i = \lambda \sum_j^I w_{ij} y_j + \alpha + \mathbf{x}_i^T \beta + \theta \sum_j^I w_{ij} \mathbf{x}_j + \varepsilon_i$
$\theta = 0$
Spatial Lag Model
$y_i = \lambda \sum_j^I w_{ij} y_j + \alpha + \mathbf{x}_i^T \beta + \varepsilon_i$
$\rho = 0$
$\lambda = 0$
$\lambda = 0$
Spatial Durbin Error Model
$y_i = \theta \sum_j^I w_{ij} \mathbf{x}_j + \alpha + \mathbf{x}_i^T \beta + \varepsilon_i$
$\xi_i = \rho \sum_j^I w_{ij} \xi_j + \varepsilon_i$
General Nesting Spatial Model
$y_i = \lambda \sum_j^I w_{ij} y_j + \alpha + \mathbf{x}_i^T \beta + \theta \sum_j^I w_{ij} \mathbf{x}_j + \varepsilon_i$
$\xi_i = \rho \sum_j^I w_{ij} \xi_j + \varepsilon_i$
OLS
$y_i = \alpha + \mathbf{x}_i^T \beta + \varepsilon_i$
$\theta = 0$
SLX
$y_i = \theta \sum_j^I w_{ij} \mathbf{x}_j + \alpha + \mathbf{x}_i^T \beta + \varepsilon_i$
$\rho = 0$
$\lambda = 0$
$\theta = 0$
$\rho = 0$
$\rho = 0$
Spatial Error Model
$y_i = \alpha + \mathbf{x}_i^T \beta + \xi_i$
$\xi_i = \rho \sum_j^I w_{ij} \xi_j + \varepsilon_i$
SAC
$y_i = \lambda \sum_j^I w_{ij} y_j + \alpha + \mathbf{x}_i^T \beta + \varepsilon_i$
$\xi_i = \rho \sum_j^I w_{ij} \xi_j + \varepsilon_i$
$\theta = 0$
$\lambda = 0$

Figure 4.1: Relationships between the different spatial dependence models

4.2 The panel data model

The spatial dependence model is only applied to cross-sectional data. When both time and spatial variables are included in the model, panel data methodology must be introduced. Panel data comprises cross-sectional observations (countries, areas, companies, households, etc.) involving measurements over time. Time series and cross-sectional data are viewed as special cases of panel data. Generally, there are three types of panels: micro, macro, and random field panels. In micro panels, there are more cross-sectional observations than time periods. Conversely, macro panels have fewer cross-sectional observations than time periods. Random field panels have a wide temporal and transverse dimension.

Hsiao (2014), Klevmarken (1989) and Baltagi (2008) demonstrated the advantages of using panel data over time-series or cross-sectional data. Compared with time-series data, panel data reduces collinearity among variables and provides more informative data as variability is added by a cross-sectional dimension. In contrast, panel data has an advantage over cross-sectional data as it captures how variables evolve in a process (Deaton, 1995). It is useful to assess intertemporal relations and lifecycle models. Besides, individuals, provinces, or countries are assumed to be heterogeneous in panel data, unlike in time-series or cross-sectional data. Generally, independent variables vary across time and geographical units. However, some independent variables are time- and/or space-invariant, which also affects the dependent variable, a possibility not considered in time series and cross-sectional data.

A panel data model is a regression model with double subscripts on its variables, for province i in year t,

$$y_{it} = \alpha + \boldsymbol{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + u_{it}, \quad i = 1, ..., I \text{ and } t = 1, ...T \tag{4.5}$$

where y_{it} is the independent variable, α is a constant, $\boldsymbol{\beta}$ is a $K \times 1$ vector and $\boldsymbol{x}_{it}$ denotes a $K \times 1$ vector of independent variables, K is the total number of independent variables, and u_{it} is the disturbance term.

Two models can be used to describe the disturbance term. The one-way error component regression model is defined as follows:

$$u_{it} = \mu_i + \varepsilon_{it}. \tag{4.6}$$

The two-way error component regression model is represented as

$$u_{it} = \mu_i + \nu_t + \varepsilon_{it}, \tag{4.7}$$

where μ_i is the unobservable space-specific effect, ν_t denotes the unobservable time effect and ε_{it} is the remainder disturbance term. Notably, μ_i and ν_t are time-invariant and space-invariant, respectively, and they explain any corresponding space-specific effect (for the former) and time-specific effect (for the latter) that is not included in the regression. If μ_i and ν_t are assumed to be fixed parameters to be estimated, and ε_{it} is independent and identically distributed, then a fixed effects model applies. If μ_i, ν_t and ε_{it} are all independent and identically distributed respectively and independent of each other, a random effects model applies.

4.3 The spatial panel data model

The cross-sectional spatial dependence model can be extended using panel data into the spatial panel data model.

For the SAR model, by modifying the subscripts from i to it in Equation (4.2), we obtain

$$y_{it} = \lambda \sum_{j}^{I} w_{ij} y_{jt} + \alpha + \boldsymbol{x}_{it}^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_{it}. \tag{4.8}$$

The SEM model can be extended in the same way using Equation (4.4).

$$\begin{aligned} y_{it} &= \alpha + \boldsymbol{x}_{it}^{\mathsf{T}} \boldsymbol{\beta} + \xi_{it} \\ \xi_{it} &= \rho \sum_{j}^{I} w_{ij} \xi_{jt} + \varepsilon_{it}, \end{aligned} \tag{4.9}$$

Compared with time-series and cross-sectional data, panel data can control for spatial and temporal heterogeneity, which is not reflected in Equations (4.8) and (4.9). Thus, the two-way error component structure is applied to address this issue. To do so, the space-specific effect μ_i and the time-specific effect ν_t are added to measure all time-invariant and space-invariant variables that may influence the estimation. Therefore, an extension of the SAR model is considered as follows:

$$y_{it} = \alpha + \lambda \sum_{j=1}^{I} w_{ij} y_{jt} + \boldsymbol{x}_{it}^{\mathsf{T}} \boldsymbol{\beta} + \mu_i + \nu_t + \varepsilon_{it}, \tag{4.10}$$

An extension of the SEM model is also considered:

$$\begin{aligned} y_{it} &= \alpha + \boldsymbol{x}_{it}^{\mathsf{T}}\boldsymbol{\beta} + \mu_i + \nu_t + \xi_{it} \\ \xi_{it} &= \rho \sum_{j}^{I} w_{ij}\xi_{jt} + \varepsilon_{it}, \end{aligned} \tag{4.11}$$

where,

i is the province and t represents the year;

y_{it} is the dependent variable;

$\boldsymbol{x}_{it}$ is the $K \times 1$ vector of explanatory variables;

K is the total number of independent variables;

w_{ij} is the element of the spatial weight matrix W for province i and j;

α is a constant parameter to be estimated;

$\boldsymbol{\beta}$ is a $K \times 1$ vector of unknown parameters to be estimated;

ε_{it} is the disturbance term and is assumed to independently and identically distributed with zero mean and constant variance;

μ_i and ν_t are the unobservable space-specific effects and time-specific effects, respectively (either fixed or random effects) defined in Equation (4.7);

λ is the spatial autoregressive coefficient;

ρ is the spatial autocorrelation coefficient.

4.4 Multicollinearity in the regression model

The regression model analyses the relationship between the dependent variable $\boldsymbol{Y}$ and independent variables $\boldsymbol{X_1, X_2, ...X_k}$. Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. This naturally means that the chance of having multicollinearity is relatively high when too many independent variables are involved in the regression model. The existence of multicollinearity will affect the estimation of the model and the explanation of the results. In the meantime, multicollinearity will lead to skewed or even incorrect interpretation (Silvey, 1969). The spatial panel data model is one particular regression model, which means multicollinearity analysis is necessary in the study.

In the first step, it is necessary to measure the presence of multicollinearity in the model. The red indicator (Kovács et al., 2005) can be used to measure average multicollinearity in the database,

$$\text{Red} = \sqrt{\frac{\sum\limits_{j=1}^{I}\sum\limits_{j=1;i\neq j}^{I} r_{ij}^{2}}{I(I-1)}},$$

where r_{ij} is the correlation coefficient of the $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. In particular, a zero red indicator implies no multicollinearity and the multicollinearity is at maximum when the red indicator equals one.

Apart from the red indicator, the condition number also measures the presence of multicollinearity through the eigenvalues of the correlation matrix of the regressors (Belsley et al., 2005). The condition number (CN) is

$$CN = \left(\frac{\lambda_{max}}{\lambda_{min}}\right)^{\frac{1}{2}}$$

where λ_{max} is the largest eigenvalue and λ_{min} is the smallest eigenvalue. Theoretically, a zero eigenvalue indicates exact collinearity and the closer it is to zero the higher degree the multicollinearity has. In practice, the moderate multicollinearity is believed to exist if the condition number is larger than 10 and the multicollinearity is strong if CN is over 30. (Belsley, 1982; Kim, 2019).

Besides, the determinant of the correlation matrix (Cooley and Lohnes, 1971), Farrar test of chi-square (Farrar and Glauber, 1967) and Theil's indicator (Theil, 1971) can be used to measure the presence of multicollinearity.

When there exist multicollinearity is in a regression model, the next step is to find multicollinearity for each independent variable in the model. The most frequently used method is the variance inflation factor (Marquaridt, 1970). For the kth independent variable, the VIF is defined as

$$\text{VIF} = \frac{1}{(1 - R_k^2)},$$

where R_k^2 is the multiple R^2 for the regression of $\boldsymbol{X}_k$ on the other covariates. If all variables are uncorrelated with each other, their VIFs should be equal to one. The higher the value of VIF, the higher the correlation between the variables. Conventionally, when the VIF is larger than 10, the variable should be removed from the regression model because of multicollinearity. A VIF value between 5 and 10 represents a high correlation that may lead to inaccurate estimation (Belsley, 1991; Gareth et al., 2013; Akinwande et al., 2015).

The Leamer's method (Greene, 2002) can also be used to check multicollinearity. Leamer proposed the method based on the variance of estimated coefficients,

$$c_k = \left\{ \frac{(\sum_i (X_{ik} - \bar{X}_k)^2)^{-1}}{(\boldsymbol{X'X})_{kk}^{-1}} \right\}^{\frac{1}{2}},$$

where $(\boldsymbol{X'X})_{kk}^{-1}$ is the kkth element of the matrix $(\boldsymbol{X'X})^{-1}$. If c_k is equal to one, the variable is uncorrelated with other variables. When c_k is close to zero, a high multicollinearity exists.

Meanwhile, Farrar and Glauber (Farrar and Glauber, 1967) used the F-test to determine the collinear regressors, where the null hypothesis is $R_i^2 = 0$. Then we have

$$\Gamma_i = \frac{R_i^2}{1 - R_i^2} \left(\frac{M - K}{K - 1} \right)$$

where K is the number of independent variables, M is the sample size, R_i^2 is the multiple correlation coefficient between $\boldsymbol{X}_i$ and the other independent variables. The random variable Γ_i follows the F-distribution with $M - K$ and $K - 1$ degrees of freedom, which is interpreted as the ratio of explained to unexplained variance. In the case that $\Gamma_i > F$, the variable $\boldsymbol{X}_i$ has multicollinearity.

Theoretically, it is ideal that there does not exist multicollinearity among independent variables in a regression model. However, it is not easy to realise in practice. How to reduce multicollinearity to a reasonable level and make sure there is no influence in the model is very important. Table 4.1 shows the summary of the above methods.

Method	Range	Multicollinearity
The presence of multicollinearity		
Red indicator	[0, 1]	The larger, the stronger multicollinearity
Condition number	≥ 0	$\geq$ 10 there is moderate multicollinearity $\geq$ 30, there is significant multicollinearity
Multicollinearity in independent variable		
VIF	≥ 1	The larger, the stronger multicollinearity $\geq$ 5, there is multicollinearity $\geq$ 10, there is strong multicollinearity
Leamer's method	[0, 1]	The smaller, the stronger multicollinearity
Farrar&Glauber test	≥ 1	the value $\Gamma_i > F$, then the independent variable has multicollinearity

Table 4.1: Summary of multicollinearity methods

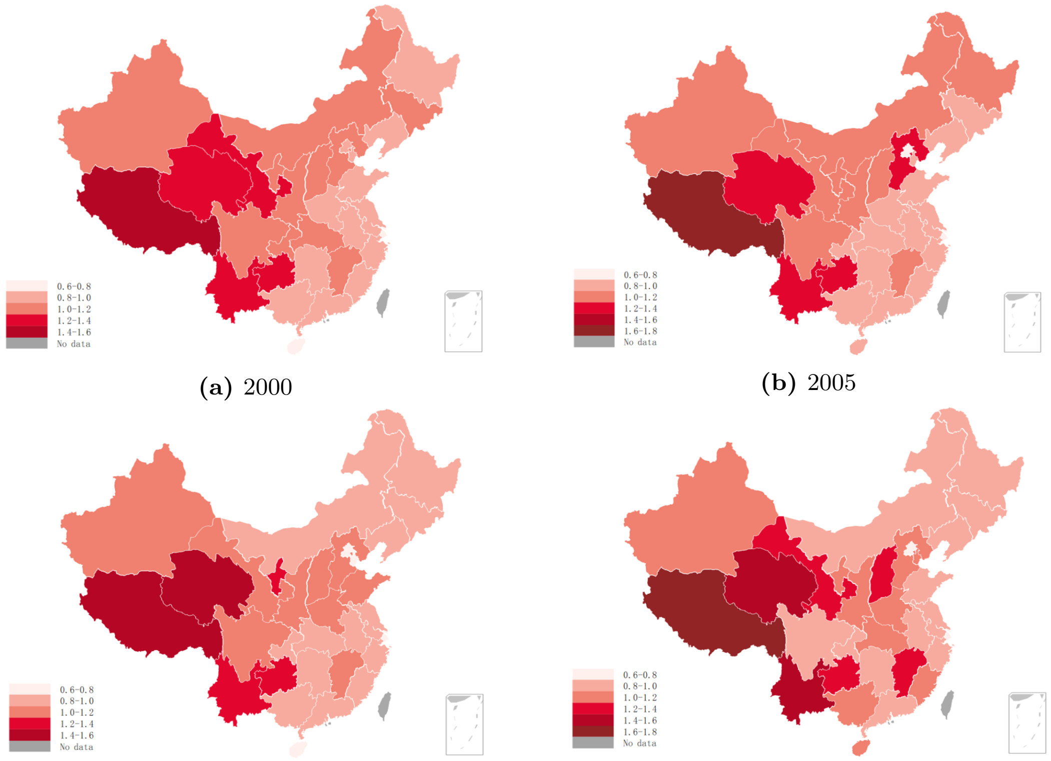
Chapter 5

Results

This chapter presents the results of this study, including heat maps, Moran scatter plots, and local indicators of spatial association (LISA) maps for the SMR at the provincial level for China in 2000, 2005, 2010, and 2015. The table of local Moran's I values for 2000 is also presented here. The tables and figures for the other years are available but have not been shown in this thesis. The spatial analysis is based on R programming. The R-packages splm (Millo et al., 2012), plm (Croissant and Millo, 2008), spdep (Bivand et al., 2005), car (Fox et al., 2007) and mctest (Ullah et al., 2019)were used in this study.

5.1 The spatial autocorrelation of mortality

The SMR was chosen to analyse all age-specific mortality rates at the provincial level for China. Figure 5.1 shows the heat maps of the SMR in 2000, 2005, 2010, and 2015. The provinces with the darker red colour have higher SMR values. A significant difference can be observed between the coastal eastern and western provinces between 2000 and 2015. Specifically, the eastern coastal provinces centred around Shanghai (marked 9 in the map), such as Jiangsu (10) and Zhejiang (11), have low SMR values. The unbalanced structure between eastern and western China has existed for many years. Geographically, western China comprises mountains, plateaus, and deserts, whereas eastern China comprises plains and low hills. Western provinces, such as Tibet (26) and Qinghai (29), experience adverse terrain and climate, which make it difficult to build roads or carry out agriculture. The low-oxygen environment in high-altitude regions simultaneously increases mortality and reduces life expectancy (Niermeyer et al., 2009). Eastern China has a better climate and more suitable topography for farming. In the modern era, south-eastern China, which is coastal, has had more opportunities to develop external networks and trade (Wan, 2008). These provinces have thus developed more rapidly, offering better healthcare and leading to lower mortality rates. In contrast, western China has relatively poorer healthcare and educational infrastructure, which affects the SMR significantly (Gyaltsen et al., 2014). Notably, the SMRs for provinces in north-eastern China have reduced since 2010.



(a) 2000

(b) 2005

(c) 2010

(d) 2015

Figure 5.1: SMR maps for China

Figure 5.1 shows an evident spatial cluster characteristic, especially in eastern and western China. Therefore, the Global Moran's I was applied to confirm the presence of spatial autocorrelation. Table 5.1 presents the values of the global Moran's I and their respective p-values. Between 2000 and 2015, the values were positive for China, with p-values less than 0.05. When the SMR in a province increased/decreased, its neighbours' SMR also increased/decreased. Apart from a slight increase in 2001 and 2006, the global Moran's I continuously decreased, from 0.401 to 0.242, in the 2001–2015 period. Thus, the spatial autocorrelation of the SMR has become weaker at the provincial level in China.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Moran's I	0.395	0.401	0.399	0.392	0.379	0.363	0.365	0.365
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Year	2008	2009	2010	2011	2012	2013	2014	2015
Moran's I	0.361	0.353	0.342	0.331	0.314	0.292	0.267	0.242
P-value	0.000	0.000	0.001	0.001	0.001	0.002	0.005	0.008

Table 5.1: Global Moran's I values of the SMR at the provincial level in China

As spatial autocorrelation was present, the local Moran's I was applied to assess the influence of individual locations and to identify cluster and outlier values. Table 5.2 shows the values of the local Moran's I at the provincial level in China and their respective p-values for the year 2000. A province with a positive local Moran's I belongs to either HH or LL, but further information (stated below) is required to determine which of the two it belongs to. Such a province is said to be a spatial cluster if its p-value is smaller than the significant level. In contrast, a province with a negative local Moran's I belongs to either HL or LH, and it is called spatial outlier if it is significant.

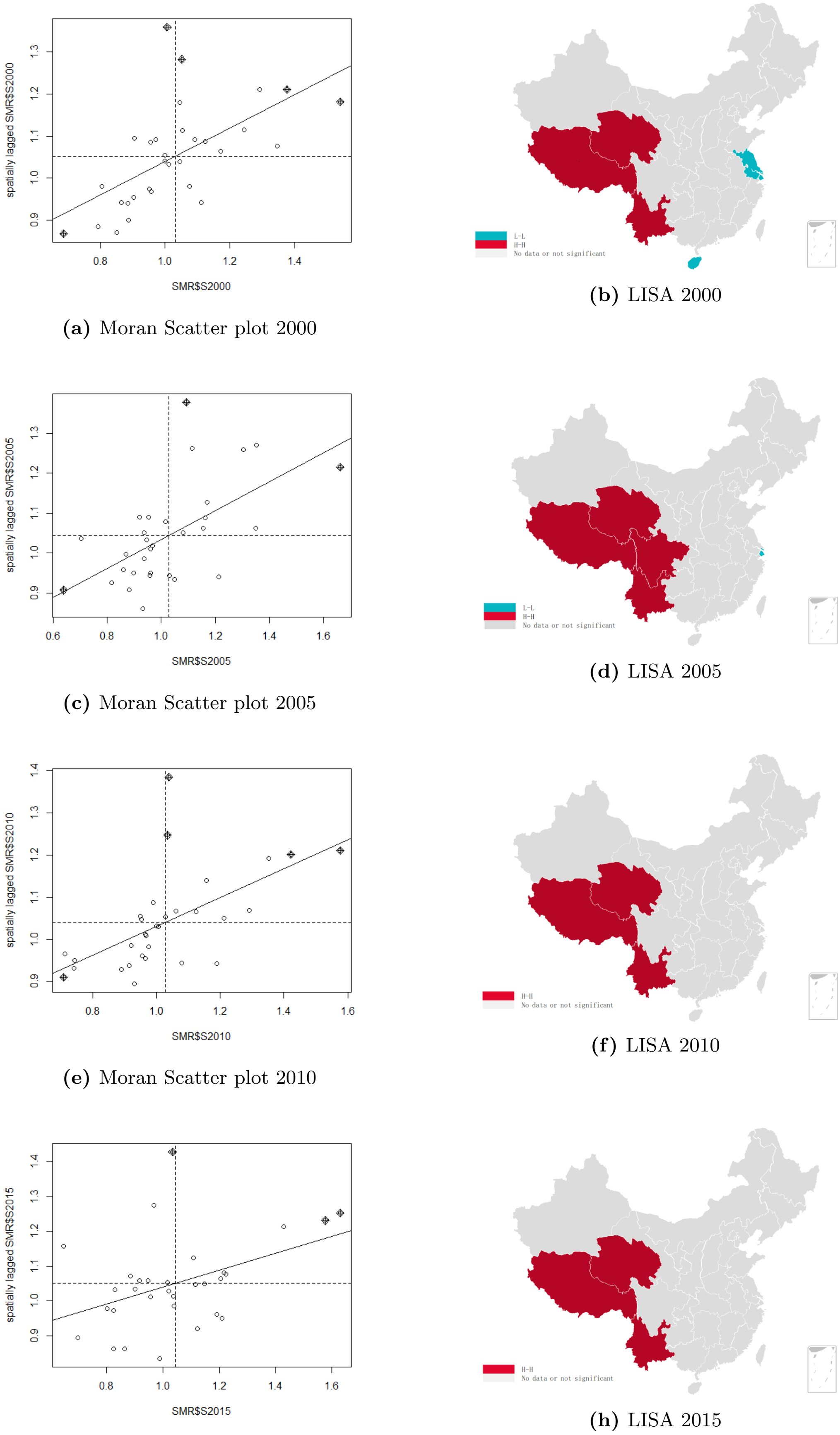
Figure 5.2 shows the Moran scatter plots and LISA maps of the SMR from 2000 to 2015. The Moran scatter plot can be considered a linear regression with the slope representing the global Moran's I value (Anselin, 1996). The abscissa represents the SMR, and the ordinate is the spatially lagged SMR (W·SMR). This graph reflects the relationship between the SMR in one province and the average SMR of its neighbours. The vertical and horizontal dotted lines indicate the average of the SMR and W·SMR respectively, dividing all provinces into four quadrants. A province located in the upper right (HH) or the lower left (LL) has spatial association with similar SMRs, implying that the province is surrounded by neighbours with similar SMRs. In contrast, the upper left (LH) and lower right (HL) quadrant provinces have spatial association with dissimilar SMRs, implying that

	Province	Local Moran's I	P-value
1	Beijing	0.358	0.278
2	Tianjin	0.410	0.252
3	Hebei	-0.067	0.541
4	Shanxi	0.109	0.377
5	Inner Mongolia	0.137	0.284
6	Liaoning	-0.126	0.569
7	Jilin	0.000	0.475
8	Heilongjiang	-0.108	0.545
9	**Shanghai**	**1.722**	**0.004**
10	**Jiangsu**	**0.895**	**0.021**
11	Zhejiang	0.583	0.062
12	Anhui	0.140	0.314
13	Fujian	0.306	0.262
14	Jiangxi	-0.220	0.699
15	Shandong	0.144	0.348
16	Henan	-0.009	0.473
17	Hubei	0.003	0.460
18	Hunan	-0.023	0.488
19	Guangdong	0.459	0.109
20	Guangxi	-0.240	0.698
21	**Hainan**	**1.077**	**0.047**
22	Chongqing	0.051	0.417
23	Sichuan	0.146	0.290
24	Guizhou	0.428	0.124
25	**Yunnan**	**1.875**	**0.000**
26	**Tibet**	**2.312**	**0.000**
27	Shaanxi	0.153	0.266
28	Gansu	0.530	0.058
29	**Qinghai**	**1.407**	**0.001**
30	Ningxia	0.059	0.431
31	Xinjiang	-0.259	0.664

Table 5.2: Local Moran's I values of the SMR at the provincial level in China in 2000

their SMRs differ considerably with their neighbours'. Using the scatter plots, it can be determined whether a province belongs to HH or LL. This coincides with the local Moran's I results.

The LISA map shows the provinces with significant values using local Moran's I values. There were two kinds of spatial clusters, HH and LL, in 2000 and 2005. Shanghai (9), Jiangsu (10), and Hainan (21) were found to be associated with the LL spatial cluster in 2000 (marked in bold in Table 5.2). However, the LL spatial cluster only appeared for Shanghai (9) in 2005 and disappeared after 2010. The HH spatial cluster was mainly distributed across Yunnan (25), Tibet (26), and Qinghai (29) between 2000 and 2015. These three provinces and their neighbours had relatively high SMRs and slow improvements in healthcare conditions during 2000–2015.



(a) Moran Scatter plot 2000

(b) LISA 2000

(c) Moran Scatter plot 2005

(d) LISA 2005

(e) Moran Scatter plot 2010

(f) LISA 2010

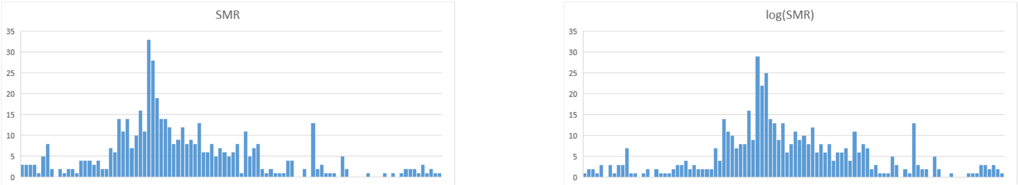
(g) Moran Scatter plot 2015

(h) LISA 2015

Figure 5.2: Moran scatter plot and LISA map of the SMR

5.2 Spatial lag model with fixed effects

Given the presence of spatial autocorrelation of mortality, the spatial panel data model was then used in this study. As the dependent variable (the SMR) showed a clear asymmetry (see Figure 5.3a), logarithmic transformation was used to convert the regressor to its log scale (see Figure 5.3b). Specifically, the SMR had a skewness of 0.871, whereas log(SMR) had a reduced skewness of 0.223.



(a) SMR **(b)** log(SMR)

Figure 5.3: Asymmetry of SMR and the log(SMR) representation

5.2.1 Multicollinearity analysis

This study considers nine factors (the details shown in Section A.2.2) as independent variables, which across four aspects, including demography, environment, economy, and societal development. Since multicollinearity is very common when many variables are considered in the model, it is necessary to check whether there exists multicollinearity among variables. One way to measure the presence of overall multicollinearity is to consider applying the condition number. The value of the condition number in this study is 80.586 which is much larger than 30. This indicates that there is a strong multicollinearity (Kim, 2019). Hence it is necessary to reduce multicollinearity among variables before creating the spatial panel data model. In this study, independent variables with VIF values less than 5 were selected. Table 5.3 illustrates the choice of independent variables. Firstly, VIFs were calculated for all independent variables, and the crude birth rate was eliminated as it has the largest VIF (larger than 5). Secondly, the GRP per capita was removed, followed by urbanisation in the third step. Eventually, the remaining independent variables have VIFs less than 5.

		Value of VIF			
1	GRP per capita	28.80	28.17 (Remove)		
2	CrudeBirthRate	33.74 (Remove)			
3	HouseholdConsumption	25.52	25.36	4.47	2.96
4	ParkGreen	1.71	1.71	1.64	1.64
5	SO2	2.03	2.02	2.00	1.99
6	DependencRatio	3.96	3.21	3.10	2.66
7	PopulationGrowth	25.57	2.33	2.19	2.02
8	ElectricalConsumption	3.59	3.57	3.56	3.53
9	Urbanisation	6.15	5.59	5.30 (Remove)	

Table 5.3: The selection of independence variables when using log(SMR) as the dependent variable using VIF

Although the Farrar & Glauber test and Leamer's method only give a range from low to high multicollinearity, they can reflect which independent variable has the highest multicollinearity. Table 5.4 shows the independent variables with the top three high multicollinearities using Farrar & Glauber test (F-G tests in table) and Leamer's method (Leamer in table). In the first method, the independent variable with a higher value shows higher multicollinearity. On the contrary, the independent variable with a higher value in Leamer's method has lower multicollinearity. The result shows that the multicollinearity analysis with different methods leads to the same result. To conclude, the independent variables with the top three high multicollinearities are crude birth rate, GRP per capita and urbanisation with regards to VIF, Farrar & Glauber test and Leamer's method. After these independent variables are removed, the condition number becomes 29.58, which is lower than 30, indicating that multicollinearity is a reasonable level.

		F-G test			Leamer		
1	GRP per capita	1938	**2214**		0.19	**0.19**	
2	CrudeBirthRate	**2283**			**0.17**		
3	HouseholdConsumption	1709	1985	340	0.20	0.20	0.47
4	ParkGreen	50	58	63	0.76	0.77	0.78
5	SO2	72	83	98	0.70	0.70	0.71
6	DependencRatio	206	180	206	0.50	0.56	0.57
7	PopulationGrowth	1713	108	117	0.20	0.66	0.68
8	ElectricalConsumption	180	210	251	0.53	0.53	0.53
9	Urbanisation	359	374	**421**	0.40	0.42	**0.43**

Table 5.4: The independent variables with the top three high multicollinearities using Farrar & Glauber test and Leamer's method

5.2.2 Model selection

In Section 5.1, the SMR maps show that the SMR has a spatial correlation at the provincial level of China. At the same time, the existence of SMR spatial correlation in China has been proven by the global and local Moran index. The SMR in China has the positive autocorrelation and the SMR clusters are located in eastern and western China. According to Elhorst (2014), the spatial dependence model can be used to explain the behaviour of SMR in provinces. However, there are space and time-specific variables affecting the SMR since this study is based on 31 provinces in 16 years. In order to control space and time-specific effects, the panel data model is applied and then we considered the spatial panel data model instead of the spatial dependence model. To determine which model best fits the data, the statistic of panel data is introduced firstly. Table 5.5 reports the statistic results when only the panel data model was considered. The table includes the OLS model, fixed effect model, and random effects model. The fixed effects model is a suitable specification when the research focus is on a specific set of observations. If the set of observations is large, the fixed effects model leads to a loss in the degrees of freedom. Therefore, the random effects model is considered to be more appropriate because it uses a subset of individuals to represent the whole population (Baltagi, 2008). Meanwhile, the fixed effects model and random effects model have included three types of effects: space-specific effects, time-specific effects, and both (space-specific and time-specific) effects. It is found that the time-specific model best fits our data, since the adjust R^2 is the closest to one among all three. At the same time, the Hausman test (Hausman, 1978) can be used to determine whether the effect is fixed or random. The null hypothesis is that the preferred model has random effects and the alternative hypothesis states that the model has fixed effects. In this study, the degree of freedom is equal to 6 because there are six independent variables in the model. The Hausman statistic was equal to 123.68, and its corresponding p-value is close to zero (much smaller than 0.05); the null hypothesis is thus rejected, indicating that the model has fixed effects.

Effect	Model	Total Sum of Squares	Residual Sum of Squares	R^2	Adjust R^2
-	OLS	3.1862	1.7876	0.4390	0.4321
Fixed	Space-specific	0.2092	0.2016	0.0360	-0.0396
	Time-specific	3.1849	1.2313	0.6134	0.5963
	Both	0.2079	0.1980	0.0477	-0.0616
Random	Space-specific	0.2479	0.2404	0.0302	0.0183
	Time-specific	3.1862	1.7876	0.4390	0.4321
	Both	0.2484	0.2409	0.0303	0.0184

Table 5.5: Statistic results of dependent variable log(SMR) using panel data models

To determine which type of spatial panel data model best fits the data, the Lagrange Multiple (LM) test and robust LM test can be applied (Burridge, 1980; Anselin, 2013). When the LM test is used for the spatial lag model, the null hypothesis is that there is no spatially lagged dependent variable. Similarly, if the LM test is used for the spatial error model, the null hypothesis assumes that there are no spatially autocorrelated error terms. If the LM test is significant in the spatial lag model, but is not significant in the spatial error model, the spatial lag model should be applied. In the converse case, the spatial error model will be the choice. When the LM tests for both models are significant, robust LM tests should be used to select the appropriate model. The robust LM test is a test for a spatial lag dependent variable in the presence of spatial error dependence (when it is a spatial lag model) and for spatial error dependence in the presence of a spatial lag dependent variable (when it is a spatial error model) (Anselin et al., 1996). If robust LM tests are significant in both models but one is magnitude significant in the other, such as p <0.00001 compared with p <0.03, then the model with more significance is selected (Anselin et al., 2008). If robust LM tests are highly significant in both models, LeSage and Pace (2009) suggested that the spatial Durbin model should be considered. Table 5.6 summarizes the selection of models using LM and robust LM tests. LM(lag) and LM(error) are LM tests for the spatial lag model and spatial error model. LM_{rob}(lag) and LM_{rob}(error) are robust LM tests for spatial lag model and spatial error model. The result Y indicates the test is significant, N means the test is not significant and Y(higher) is magnitude significant.

	Significant						
LM(lag)	Y	N	Y	Y	Y	Y	Y
LM(error)	N	Y	Y	Y	Y	Y	Y
LM_{rob}(lag)	-	-	Y	N	Y(higher)	Y	Y
LM_{rob}(error)	-	-	N	Y	Y	Y(higher)	Y
Model selection	lag	error	lag	error	lag	error	durbin

Table 5.6: Model selection in LM and robust LM test

Table 5.7 shows the results of spatial panel data model selection. It is used LM test in both spatial lag model and spatial error model with log(SMR) and log(CMF) as the dependent variable. Moreover, three kinds of fixed effects are considered. The table shows that the null hypothesis was not rejected at the 5%

level of significance in all cases. The null hypothesis was rejected at the 10% level of significance only for the spatial lag model with time-specific effects. Thus, the spatial lag model with time-specific effects was applied in this study with log(SMR) as the dependent variable.

Dependent variable	Test	Space-specific effect μ_i		Time-specific effects ν_t		Spatial-specific and time specific effects	
		Value	p-value	Value	p-value	Value	p-value
log(SMR)	LM(lag)	0.229	0.633	3.275	0.070	0.124	0.725
	LM(error)	0.241	0.624	0.673	0.412	0.049	0.826
log(CMF)	LM(lag)	0.193	0.661	1.158	0.282	0.139	0.709
	LM(error)	0.141	0.708	1.337	0.248	0.045	0.832

Table 5.7: The selection of spatial panel data model

5.2.3 Model estimation

Finally, ML estimation was used to estimate all unknown parameters. The estimation results are presented in Table 5.8. As shown in the table, the parameters of the independent variables SO_2 and dependency ratio are not significant. Hence, these two variables were removed from the model.

	Estimate	Std.Error	P-value	
α	0.2142	0.0533	0.0000	***
λ	0.102438	0.048222	0.033640	*
$\beta_{\text{Household consumption}}$	-0.021080	0.001398	0.000000	***
$\beta_{\text{Park green}}$	-0.019262	0.002446	0.000000	***
β_{SO2}	0.000002	0.000585	0.997700	
$\beta_{\text{Dependency ratio}}$	0.000521	0.001240	0.674200	
$\beta_{\text{Population grown}}$	0.018005	0.002461	0.000000	***
$\beta_{\text{Electrial consumption}}$	0.018275	0.004580	0.000066	***

$***$: p-value<0.000, $*$: p-value<0.05.

Table 5.8: Estimation results for the spatial lag model with fixed effects

ML estimation was applied once again after removing the two variables. Table 5.9 presents the updated estimation results of the spatial lag model with fixed effects. The mean intercept α was obtained from $\bar{y}_{..} - \beta\bar{x}_{..}$ (Baltagi, 2008), where $\bar{y}_{..}$ and $\bar{x}_{..}$ denoted the average values of y_{it} and x_{it} for all provinces i and year t, respectively. The spatial autoregressive coefficient λ was 0.1022, which was significant in the model with p-value equal to 0.0339. Thus, neighbours influence each province. The household consumption and park greenery level had a negative relationship with the log(SMR). A 1 yuan increase in household consumption would lead to a 0.0213 decrease in the log(SMR). These findings are similar to those of Schultz (1984) and Kinge et al. (2019). Furthermore, a 1 m^2 per capita increase in park green area would result in a 0.0193 decrease in the log(SMR). This relationship coincides with the results of Alcock et al. (2014), Twohig-Bennett and Jones (2018), Kondo et al. (2018) and Wang and Tassinary (2019), who proved that the improvement of green area in a residential environment can reduce mortality directly or indirectly.

	Estimate	Std.Error	P-value	
α	0.2333	0.0279	0.0000	***
λ	0.1022	0.0482	0.0339	*
$\beta_{\text{Household consumption}}$	-0.0213	0.0012	0.0000	***
$\beta_{\text{Park green}}$	-0.0193	0.0024	0.0000	***
$\beta_{\text{Population grown}}$	0.0186	0.0019	0.0000	***
$\beta_{\text{Electrial consumption}}$	0.0179	0.0032	0.0000	***

$***$: p-value<0.000, $*$: p-value<0.05.

Table 5.9: Estimation results of the spatial lag model with fixed effects

This study also found that population growth and electrical consumption were positively associated with the log(SMR). If the population growth rate increased by 1, the log(SMR) would increase by 0.0186. Similarly, a 1,000 kwh per capita increase in electrical consumption would lead to a 0.00179 increase in the log(SMR). Notably, electrical consumption as used in this study is the combined industrial and residual consumption. Although Mazur (2011) mentioned that an increase in residual electricity consumption is essential for the improved well-being of poor areas and countries, industrial electrical consumption comprises the lion's share of electrical consumption (85.94% in 2018[1]) in China and generates air pollution. Therefore, electrical consumption has a positive relationship with the log(SMR) for China. Gohlke et al. (2011) pointed out that while there is no significant relationship between electrical consumption and life expectancy, the increase in electrical consumption in China lead to the increase in infant deaths.

Table 5.10 shows the results of the estimation of time-specific effects ν_t in the spatial lag model with fixed effects, which includes the estimated values of ν_t, SEs and the p-values. The term ν_t represents the deviation from α in year t and can be obtained from $\bar{y}_{.t} - \beta\bar{x}_{.t} - \alpha$ (Baltagi, 2008), where $\bar{y}_{.t}$ and $\bar{x}_{.t}$ denote the average of $\sum_i y_{it}$ and $\sum_i x_{it}$ for all years respectively. The negative sign of ν_t shows that the average of the SMR for this year is lower than the overall average SMR.

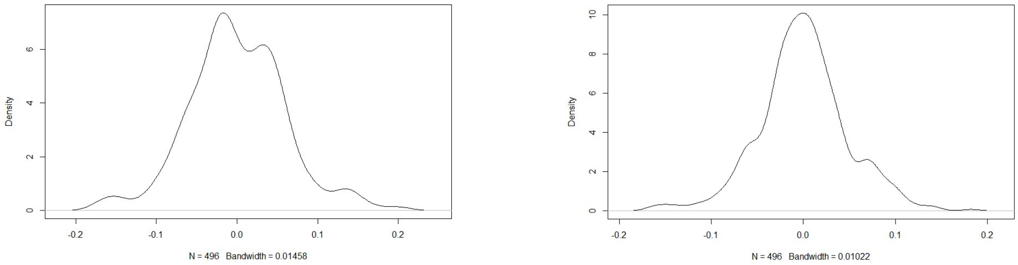
[1] Data source: the NBS

	Estimate ν_t	Std.Error	P-value	
2000	-0.1579	0.0296	0.0000	***
2001	-0.1404	0.0299	0.0000	***
2002	-0.1660	0.0275	0.0000	***
2003	-0.1383	0.0281	0.0000	***
2004	-0.1181	0.0293	0.0001	***
2005	-0.1009	0.0302	0.0008	***
2006	-0.0791	0.0313	0.0114	*
2007	-0.0523	0.0327	0.1102	
2008	-0.0176	0.0338	0.6021	
2009	0.0137	0.0356	0.7003	
2010	0.0491	0.0370	0.1850	
2011	0.1039	0.0390	0.0077	**
2012	0.1372	0.0404	0.0007	***
2013	0.1799	0.0418	0.0000	***
2014	0.2191	0.0437	0.0000	***
2015	0.2679	0.0447	0.0000	***

$***$: p-value<0.000, $*$: p-value<0.05.

Table 5.10: Estimation results of the time-specific effects in the spatial lag model with fixed effects

It is necessary to do the goodness-of-fit tests and residual analysis for the model. The Akaike Information Criterion (AIC) (Akaike, 1998) is used to compare models and evaluate which one has the best fitting. The lower AIC, the better the model is. In the study, the spatial lag model with fixed effects has an AIC of -699.865 whereas the linear model with no lags has an AIC of -541.0281, which suggests that the first model is better. Besides, the residual density plot is used to analyse the behaviour of residuals. Figure 5.4 shows the residual density plots of OLS and the spatial lag model with fixed effects (SLMFE). It is found that the residual density plot of SLMFE is closer to bell-shape curve than the residual density plot of OLS, which the (standard) normal density possesses. At the same time, the Moran test is used in residuals of the spatial lag model with fixed effects and the results are shown in Table 5.11. It is found that all Moran's I values are closed to zero, which indicates that there is almost no spatial autocorrelation. Moreover, the



(a) The residual density plot of OLS

(b) The residual density plot of SLMFE

Figure 5.4: The residual density plot of OLS and SLMFE

null hypothesis (no spatial autocorrelation) cannot be rejected since all p-values are larger than 0.05. Therefore, both aspects can support the selection of the proposed model.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Moran's I	-0.096	-0.079	-0.028	-0.003	-0.001	-0.019	-0.051	-0.020
P-value	0.557	0.506	0.313	0.241	0.226	0.275	0.374	0.269
Year	2008	2009	2010	2011	2012	2013	2014	2015
Moran's I	-0.055	-0.087	-0.083	-0.049	-0.048	-0.061	-0.063	-0.047
P-value	0.390	0.524	0.520	0.391	0.388	0.425	0.433	0.349

Table 5.11: Global Moran's I of the residuals of the spatial lag model with fixed effects

Chapter 6

Conclusions

China has rapidly developed its economy alongside its population's life expectancy. However, unbalanced development across China has also created challenges. The more-developed provinces attract younger cohorts, which increases economic disparities and changes the age distribution of population at the provincial level.

This study used China's census (2000, 2010) and micro census (2005, 2015) data, including age-specific population and death figures at the provincial level, to calculate the SMR. Additionally, other demographic, environmental, economic, and social data for 2000–2015 were collected from the NBS. Interpolation methods were used to account for missing data for one or more years. The application of mortality standardisation showed that mortality rates exhibit a spatial imbalance at the provincial level in China. In general, provinces in south-eastern China have lower mortality rates. This is because these provinces have well-developed economies, trade opportunities, and a climate conducive to agriculture. South-eastern China also has better living, healthcare, and educational infrastructure. In contrast, provinces in western China showed higher mortality rates for 2000–2015, owing to an adverse high-altitude, low-oxygen environment, and poor medical and educational conditions.

To analyse mortality autocorrelation between provinces, the global and local Moran's indices were used to test spatial autocorrelation. The SMR had positive spatial autocorrelation at the provincial level, implying that the change in SMR for one province was usually in the same direction as that of its neighbours. The local Moran's I values indicated the presence of high–high (HH) and low–low (LL) spatial clusters in 2000 and 2005. Only a high–high spatial cluster was observed in 2010 and 2015. A cluster of high SMR (HH spatial cluster) was formed in eastern China, whereas a cluster of low SMR (LL spatial cluster) was formed in western China.

The spatial lag model with fixed effects was then applied to model the detected spatial dependence at the provincial level. The spatial autoregressive coefficient λ was found to be 0.1022, indicating that the log(SMR) of a province changed with the log(SMR) of its neighbours. Meanwhile, household consumption, park green level, population growth, and electrical consumption were found to affect the log(SMR) at the provincial level. The first and last two impact factors had negative and positive relationships with the log(SMR), respectively.

This study provides a standardised method to analyse the differences in mortality rates within China and identify the causes driving such differences between provinces. The increase in population growth leads to an increase in mortality rates. Medical treatment and healthcare systems should be improved in underdeveloped provinces, especially those in western China. Furthermore, economic development is negatively correlated with mortality rates at the national level. However, many workers from less-developed provinces migrate to more-developed provinces (or cities) for jobs and contribute to the economies of the more-developed provinces, thereby enlarging the gap between less- and more-developed provinces in China. Pensions and other social security payments in less-developed provinces as well for workers employed in different provinces should be improved. In addition, greater investments in social security, education, and healthcare in the cities/provinces surrounding the more-developed cities can attract more migrant workers to such regions, making the gap smaller. However, economic development may inevitably lead to pollution. For example, the large industrial consumption of electricity leads to air pollution. Hence, improving environmental conditions by developing much-needed green spaces in urban areas can decrease mortality rates in China. The government should consider environmental protection and pollution in the context of sustainable development.

This study is subject to certain limitations. First, data from only four years were included. Additionally, only the data for 2000 and 2010 were complete, whereas the data for 2005 and 2015 were 1% sample data. This study can be extended in future using more recent data. Second, as government plans and policies greatly affect provincial development, future researchers can include such relevant impact factors in their models, although it is difficult to quantify these plans and policies.

Bibliography

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer.

Akinwande, M. O., H. G. Dikko, A. Samson, et al. (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics 5*(07), 754.

Alcock, I., M. P. White, B. W. Wheeler, L. E. Fleming, and M. H. Depledge (2014). Longitudinal effects on mental health of moving to greener and less green urban areas. *Environmental Science & Technology 48*(2), 1247–1255.

Andridge, R. R. and R. J. Little (2010). A review of hot deck imputation for survey non-response. *International statistical review 78*(1), 40–64.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis 27*(2), 93–115.

Anselin, L. (1996). The moran scatterplot as an ESDA tool to assess local instability in spatial. *Spatial Analytical 4*, 111.

Anselin, L. (2013). *Spatial Econometrics: Methods and Models*, Volume 4. Springer Science & Business Media.

Anselin, L. and A. K. Bera (1998). Introduction to spatial econometrics. *Handbook of Applied Economic Statistics 237*.

Anselin, L., A. K. Bera, R. Florax, and M. J. Yoon (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics 26*(1), 77–104.

Anselin, L. and S. Hudak (1992). Spatial econometrics in practice: A review of software options. *Regional Science and Urban Economics 22*(3), 509–536.

Anselin, L., J. Le Gallo, and H. Jayet (2008). Spatial panel econometrics. In *The econometrics of panel data*, pp. 625–660. Springer.

Arbia, G. (2006). *Spatial Econometrics: Statistical Foundations and Applications to Regional Convergence*. Springer Science & Business Media.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Armitage, P., G. Berry, and J. N. S. Matthews (2008). *Statistical Methods in Medical Research*. John Wiley & Sons.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*. John Wiley & Sons.

Baltagi, B. H. and D. Li (2006). Prediction in the panel data model with spatial correlation: the case of liquor. *Spatial Economic Analysis 1*(2), 175–185.

Baltagi, B. H., S. H. Song, and W. Koh (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics 117*(1), 123–150.

Banister, J. and K. Hill (2004). Mortality in China 1964–2000. *Population Studies 58*(1), 55–75.

Barufi, A. M., E. Haddad, and A. Paez (2012). Infant mortality in Brazil, 1980-2000: A spatial panel data analysis. *BMC Public Health 12*(1), 181.

Batista, G. E., M. C. Monard, et al. (2002). A study of k-nearest neighbour as an imputation method. *His 87*(251-260), 48.

Baylis, K., N. D. Paulson, and G. Piras (2011). Spatial approaches to panel data in agricultural economics: a climate change application. *Journal of Agricultural and Applied Economics 43*(3), 325–338.

Bell, K. P. and N. E. Bockstael (2000). Applying the generalized-moments estimation approach to spatial problems involving micro-level data. *Review of Economics and Statistics 82*(1), 72–82.

Belsley, D. A. (1982). Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics 20*(2), 211–253.

Belsley, D. A. (1991). *Conditioning diagnostics: Collinearity and weak data in regression.* Number 519.536 B452. Wiley.

Belsley, D. A., E. Kuh, and R. E. Welsch (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, Volume 571. John Wiley & Sons.

Benjamin, B. (1968). *Health and Vital Statistics.* London: George Allen and Unwin Ltd., Ruskin House, Museum Street, WCI.

Beretta, L. and A. Santaniello (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making 16*(3), 197–208.

Bivand, R., A. Bernat, M. Carvalho, Y. Chun, C. Dormann, S. Dray, R. Halbersma, N. Lewin-Koh, J. Ma, G. Millo, et al. (2005). The spdep package. *Comprehensive R Archive Network*, 05–83.

Bongaarts, J. and S. Greenhalgh (1985). An alternative to the one-child policy in china. *Population and Development Review*, 585–617.

Breslow, N. (1987). Statistical methods in cancer research. volume ii. *The Design and Analysis of Cohort Studies*.

Breslow, N. and N. Day (1985). *The Standardized Mortality Ratio. Biostatistics: Statistics in Biomedical Public Health and Environmental Sciences.* North Holland: Elsevier Science Publishiers.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, 45–57.

Breslow, N. E. and N. Day (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases 28*(5-6), 289–303.

Breslow, N. E., N. E. Day, and E. Heseltine (1980). *Statistical Methods in Cancer Research*, Volume 1. International agency for research on cancer Lyon.

Burridge, P. (1980). On the Cliff-Ord test for spatial correlation. *Journal of the Royal Statistical Society: Series B (Methodological) 42*(1), 107–108.

Cao, J., M. S. Ho, and D. Jorgenson (2012). An integrated assessment of the economic costs and environmental benefits of pollution and carbon control. In *The Chinese Economy*, pp. 231–256. Springer.

Carracedo, P. and A. Debón (2016). Spatial statistical tools to assess mortality differences in Europe. pp. 49–74.

Cartwright, M. H., M. J. Shepperd, and Q. Song (2004). Dealing with missing software project data. In *Proceedings. 5th International Workshop on Enterprise*

Networking and Computing in Healthcare Industry (IEEE Cat. No. 03EX717), pp. 154–165. IEEE.

Chen, S.-H., H.-W. Lin, E. Bucciarelli, F. Muratore, and I. Odoardi (2017). A data mining analysis of the Chinese inland-coastal inequality. pp. 96–104.

Chen, W. and L. Zhang (2015). A reassessment of China's recent fertility (original language is Chinese). *Population Research 39*(2), 32–39.

Chen, Y., A. Bouferguene, Y. Shen, and M. Al-Hussein (2019). Difference analysis of regional population ageing from temporal and spatial perspectives: A case study in China. *Regional Studies 53*(6), 849–860.

Cliff, A. D. and J. K. Ord (1981). *Spatial Processes: Models & Applications.* Taylor & Francis.

Coale, A. J. (1981). Population trends, population policy, and population studies in China. *Population and Development Review*, 85–97.

Cooley, W. W. and P. R. Lohnes (1971). *Multivariate data analysis.* New York: Wiley.

Cressie, N. (2015). *Statistics for Spatial Data.* John Wiley & Sons.

Croissant, Y. and G. Millo (2008, 07). Panel data econometrics in R: The plm package. *Journal of Statistical Software 27*, 1–43.

Cutler, D., A. Deaton, and A. Lleras-Muney (2006). The determinants of mortality. *Journal of Economic Perspectives 20*(3), 97–120.

Dale, W. (1777). *A Supplement to Calculations of the Value of Annuities, Published for the Use of Societies Instituted for the Benefit of Age, Containing Various Illustrations of the Doctrine of Annuities, and Compleat Tables of the Value of 1£[sic] Immediate Annuity.* J. Ridley.

Day, N. (1976). A new measure of age-standardized incidence. cumulative rate tables. *Cancer Incidence in Five Continents*, 443–452.

Deaton, A. (1995). Data and econometric tools for development analysis. *Handbook of Development Economics 3*, 1785–1882.

Démurger, S., J. D. Sachs, W. T. Woo, S. Bao, G. Chang, and A. Mellinger (2002). Geography, economic policy, and regional development in China. *Asian Economic Papers 1*(1), 146–197.

Ding, D., W. Wang, J. Wu, G. Ma, X. Dai, B. Yang, T. Wang, C. Yuan, Z. Hong, H. M. de Boer, et al. (2006). Premature mortality in people with epilepsy in rural China: a prospective study. *The Lancet Neurology 5*(10), 823–827.

Doll, R. and P. Cook (1967). Summarizing indices for comparison of cancer incidence data. *International Journal of Cancer 2*(3), 269–279.

Driscoll, J. C. and A. C. Kraay (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics 80*(4), 549–560.

Druska, V. and W. C. Horrace (2004). Generalized moments estimation for spatial panel data: Indonesian rice farming. *American Journal of Agricultural Economics 86*(1), 185–198.

Du, L., X. Wang, M. Wang, and Y. Lan (2012). Analysis of mortality in chrysotile asbestos miners in China. *Journal of Huazhong University of Science and Technology [Medical Sciences] 32*(1), 135–140.

Duda, R. O., P. E. Hart, et al. (1973). *Pattern classification and scene analysis*, Volume 3. Wiley New York.

Egger, P., M. Pfaffermayr, and H. Winner (2005). An unbalanced spatial panel data approach to US state tax competition. *Economics Letters 88*(3), 329–335.

Elhorst, J. P. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review 26*(3), 244–268.

Elhorst, J. P. (2010). Applied spatial econometrics: raising the bar. *Spatial Economic Analysis 5*(1), 9–28.

Elhorst, J. P. (2014). *Spatial Econometrics: From Cross-sectional Data to Spatial Panels*, Volume 479. Springer.

Enders, C. K. (2010). *Applied missing data analysis.* Guilford press.

Farrar, D. E. and R. R. Glauber (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92–107.

Fisher, I. (1927). *The Making of Index Numbers.* Boston: Houghton Mifflin Company.

Fleiss, J. L., B. Levin, and M. C. Paik (2013). *Statistical methods for rates and proportions.* John Wiley & Sons.

Fox, J., G. G. Friendly, S. Graves, R. Heiberger, G. Monette, H. Nilsson, B. Ripley, S. Weisberg, M. J. Fox, and M. Suggests (2007). The car package. *R Foundation for Statistical Computing*.

Frazier, C. and K. M. Kockelman (2005). Spatial econometric models for panel data: incorporating spatial and temporal data. *Transportation Research Record 1902*(1), 80–90.

Gächter, M. and E. Theurl (2011). Health status convergence at the local level: empirical evidence from austria. *International journal for equity in health 10*(1), 34.

Gareth, J., W. Daniela, H. Trevor, and T. Robert (2013). *An introduction to statistical learning: with applications in R.* Spinger.

Gascon, M., M. Triguero-Mas, D. Martínez, P. Dadvand, J. Forns, A. Plasència, and M. J. Nieuwenhuijsen (2015). Mental health benefits of long-term exposure to residential green and blue spaces: a systematic review. *International Journal of Environmental Research and Public Health 12*(4), 4354–4379.

General, R. (1841a). Annual report of the registrar general for England and Wales.

General, R. (1841b). Registrar general's statistical review of England and Wales.

General, R. (1853). Annual report of the registrar general for England and Wales.

General, R. (1857). Annual report of the registrar general for England and Wales.

General, R. (1883). Annual report of the registrar general for England and Wales.

General, R. (1884). Annual report of the registrar general for England and Wales.

Gohlke, J. M., R. Thomas, A. Woodward, D. Campbell-Lendrum, A. Prüss-Üstün, S. Hales, and C. J. Portier (2011). Estimating the global public health implications of electricity and coal consumption. *Environmental Health Perspectives 119*(6), 821–826.

Goudarzi, G., S. Geravandi, E. Idani, S. A. Hosseini, M. M. Baneshi, A. R. Yari, M. Vosoughi, S. Dobaradaran, S. Shirali, M. B. Marzooni, et al. (2016). An evaluation of hospital admission respiratory disease attributed to sulfur dioxide ambient concentration in Ahvaz from 2011 through 2013. *Environmental Science and Pollution Research 23*(21), 22001–22007.

Greene, W. H. (2002). *Econometric analysis.* Prentic Hall.

Gwatkins, D., S. Rutstein, K. Johnson, E. Suliman, A. Wagstaff, and A. Amouzou (2007). *Socio-economic Differences in Health, Nutrition, and Population Within Developing Countries*. Washington, DC, World Bank.

Gyaltsen, K., L. Gyal, J. D. Gipson, T. Kyi, and A. R. Pebley (2014). Reducing high maternal mortality rates in western China: a novel approach. *Reproductive Health Matters 22*(44), 164–173.

Haining, R. (1993). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press.

Hao, R. and Z. Wei (2010). Fundamental causes of inland–coastal income inequality in post-reform China. *The Annals of Regional Science 45*(1), 181–206.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251–1271.

Hoem, J. M. (1987). Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *International Statistical Review/Revue Internationale de Statistique*, 119–152.

Hongyang Cui, Lan Xu, R. L. (2013). An evaluation of data accuracy of the 2010 population census of China (original language is chinese). *Population Research 37*(1), 10–21.

Hou, B., J. Nazroo, J. Banks, and A. Marshall (2019). Are cities good for health? A study of the impacts of planned urbanization in China. *International Journal of Epidemiology 48*(4), 1083–1090.

Hsiao, C. (2014). *Analysis of Panel Data*. Number 54. Cambridge University Press.

Huang, F. and B. Browne (2017). Mortality forecasting using a modified continuous mortality investigation mortality projections model for China I: Methodology and country-level results. *Annals of Actuarial Science 11*(1), 20.

Inskip, H. (2014). Standardization methods. *Wiley StatsRef: Statistics Reference Online*.

Inskip, H., V. Beral, P. Fraser, and J. Haskey (1983). Methods for age-adjustment of rates. *Statistics in Medicine 2*(4), 455–466.

Jarup, L. (2004). Health and environment information systems for exposure and disease mapping, and risk assessment. *Environmental Health Perspectives 112*(9), 995–997.

Johnson, D. R. and R. Young (2011). Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family 73*(5), 926–945.

Julious, S. A., J. Nicholl, and S. George (2001). Why do we continue to use standardized mortality ratios for small area comparisons? *Journal of Public Health 23*(1), 40–46.

Kapoor, M., H. H. Kelejian, and I. R. Prucha (2007). Panel data models with spatially correlated error components. *Journal of Econometrics 140*(1), 97–130.

Kashyap, R., T. D. Singh, H. Rayes, J. C. O'Horo, G. Wilson, P. Bauer, and O. Gajic (2019). Association of septic shock definitions and standardized mortality ratio in a contemporary cohort of critically ill patients. *Journal of Critical Care 50*, 269–274.

Kelejian, H. H. and I. R. Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics 17*(1), 99–121.

Khan, R. R. and M. Siddiqui (2014). Review on effects of particulates: Sulfur dioxide and nitrogen dioxide on human health. *International Research Journal of Environment Sciences 3*(4), 70–3.

Kim, I., H.-K. Lim, H.-Y. Kang, and Y.-H. Khang (2020). Comparison of three small-area mortality metrics according to urbanity in Korea: the standardized mortality ratio, comparative mortality figure, and life expectancy. *Population Health Metrics 18*(1), 1–14.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology 72*(6), 558.

Kinge, J. M., J. H. Modalsli, S. Øverland, H. K. Gjessing, M. C. Tollånes, A. K. Knudsen, V. Skirbekk, B. H. Strand, S. E. Håberg, and S. E. Vollset (2019). Association of household income with life expectancy and cause-specific mortality in Norway, 2005-2015. *Jama 321*(19), 1916–1925.

Klevmarken, N. A. (1989). Panel studies: what can we learn from them? *European Economic Review 33*, 523–529.

Kondo, M. C., J. M. Fluehr, T. McKeon, and C. C. Branas (2018). Urban green space and its impact on human health. *International Journal of Environmental Research and Public Health 15*(3), 445.

Kovács, P., T. Petres, and L. Tóth (2005). A new measure of multicollinearity in linear regression models. *International statistical review 73*(3), 405–412.

Kowarik, A. and M. Templ (2016). Imputation with the r package vim. *Journal of Statistical Software 74*(7), 1–16.

Lai, D., F. Guo, and R. J. Hardy (2000). Standardized mortality ratio and life expectancy: a comparative study of chinese mortality. *International Journal of Epidemiology 29*(5), 852–855.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica 72*(6), 1899–1925.

Lee, L.-f. and J. Yu (2012). Qml estimation of spatial dynamic panel data models with time varying spatial weights matrices. *Spatial Economic Analysis 7*(1), 31–74.

LeSage, J. P. (1997). Bayesian estimation of spatial autoregressive models. *International Regional Science Review 20*(1-2), 113–129.

LeSage, J. P. and R. Pace (2009). *Introduction to Spatial Econometrics.* CRC Press Taylor & Francis Group, Boca Raton.

Li, J. S.-H., K. Q. Zhou, X. Zhu, W.-S. Chan, and F. W.-H. Chan (2019). A bayesian approach to developing a stochastic mortality model for China. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 182*(4), 1523–1560.

Li, X. (2017). Spatial panel model to assess population aging in regional China (original language is Chinese). *Contemporary Economy* (9), 128–131.

Li, X., J. Song, T. Lin, J. Dixon, G. Zhang, and H. Ye (2016). Urbanization and health in China, thinking at the national, local and individual levels. *Environmental Health 15*(1), 113–123.

Lilienfeld, D. E. (1978). The greening of epidemiology: Sanitary physicians and the london epidemiological society (1830-1870). *Bulletin of the History of Medicine 52*(4), 503–528.

Lin, C.-M., C.-Y. Li, G.-Y. Yang, and I.-F. Mao (2004). Association between maternal exposure to elevated ambient sulfur dioxide during pregnancy and term low birth weight. *Environmental Research 96*(1), 41–50.

Lin, W., M. Wu, Y. Zhang, R. Zeng, X. Zheng, L. Shao, L. Zhao, S. Li, and Y. Tang (2018). Regional differences of urbanization in China and its driving factors. *Science China Earth Sciences 61*(6), 778–791.

Linden, M. and D. Ray (2017). Aggregation bias-correcting approach to the health–income relationship: Life expectancy and gdp per capita in 148 countries, 1970–2010. *Economic Modelling 61*, 126–136.

Liu, Y., Y. Zheng, J. Chen, Y. Shi, L. Shan, S. Wang, W. Wang, X. Shen, and Y. Zhang (2018). Tuberculosis-associated mortality and its risk factors in a district of Shanghai, China: a retrospective cohort study. *The International Journal of Tuberculosis and Lung Disease 22*(6), 655–660.

Lleras-Muney, A. (2005). The relationship between education and adult mortality in the United States. *The Review of Economic Studies 72*(1), 189–221.

Logan, W. (1982). Cancer mortality by occupation and social class 1851-1971. *IARC Scientific Publications* (36), 1.

Lu, Q., K. Hanewald, and X. Wang (2019). Bayesian hierarchical multi-population mortality modelling for China's provinces. *ARC Centre of Excellence in Population Ageing Research (CEPAR) Working Paper* (2019/17).

Luo, Q., M. Zhang, W. Yao, Y. Fu, H. Wei, Y. Tao, J. Liu, and H. Yao (2018). A spatio-temporal pattern and socio-economic factors analysis of improved sanitation in China, 2006–2015. *International Journal of Environmental Research and Public Health 15*(11), 2510.

Luo, Y., Z. Zhang, and D. Gu (2015). Education and mortality among older adults in China. *Social Science & Medicine 127*, 134–142.

Maltamo, M., J. Malinen, A. Kangas, S. Härkönen, and A.-M. Pasanen (2003). Most similar neighbour-based stand variable estimation for use in inventory by compartments in finland. *Forestry 76*(4), 449–464.

Marquaridt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics 12*(3), 591–612.

Mazur, A. (2011). Does increasing energy or electricity consumption improve quality of life in industrial nations? *Energy Policy 39*(5), 2568–2572.

Meliker, J. R., R. L. Wahl, L. L. Cameron, and J. O. Nriagu (2007). Arsenic in drinking water and cerebrovascular disease, diabetes mellitus, and kidney disease in michigan: a standardized mortality ratio analysis. *Environmental Health 6*(1), 4.

Miettinen, O. S. (1972). Standardization of risk ratios. *American Journal of Epidemiology 96*(6), 383–388.

Millo, G., G. Piras, et al. (2012). splm: Spatial panel data models in r. *Journal of Statistical Software 47*(1), 1–38.

Mok, C., C. Kwok, L. Ho, P. Chan, and S. Yip (2011). Life expectancy, standardized mortality ratios, and causes of death in six rheumatic diseases in Hong Kong, China. *Arthritis & Rheumatism 63*(5), 1182–1189.

Mok, C., R. C. Kwok, and P. S. Yip (2013). Effect of renal disease on the standardized mortality ratio and life expectancy of patients with systemic lupus erythematosus. *Arthritis & Rheumatism 65*(8), 2154–2160.

Moran, P. A. (1950a). Notes on continuous stochastic phenomena. *Biometrika 37*(1/2), 17–23.

Moran, P. A. (1950b). A test for the serial independence of residuals. *Biometrika 37*(1/2), 178–181.

Muller, A. (2002). Education, income inequality, and mortality: a multiple regression analysis. *British Medical Journal 324*(7328), 23.

Mutl, J. and M. Pfaffermayr (2011). The hausman test in a cliff and ord panel model. *The Econometrics Journal 14*(1), 48–76.

Nations, U. (2019). World population prospects 2019: highlights. *Department of Economic and Social Affairs, Population Division.*

Niermeyer, S., P. A. Mollinedo, and L. Huicho (2009). Child health and living at high altitude. *Archives of Disease in Childhood 94*(10), 806–811.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association 70*(349), 120–126.

Pfaffermayr, M. (2009). Maximum likelihood estimation of a general unbalanced spatial random effects model: A Monte Carlo study. *Spatial Economic Analysis 4*(4), 467–483.

Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies 29*(2), 231–248.

Preston, S. H. (1980). Causes and consequences of mortality declines in less developed countries during the twentieth century. pp. 289–360.

Ren, Q., Y. You, X. Zheng, X. Song, and G. Cheng (2004). The differences of mortality level in regional China since the 1980s (original language is Chinese). *Chinese Population Science* (3), 19–29.

Rubin, D. B. (1976). Inference and missing data. *Biometrics 63*(3), 581–592.

Rubin, M. and H. Westergaard (1886). Landbefolkningens dodelighed i fyens stift (mortality of the rural population in the diocese of funen). *Copenhagen: PG Philipsen*.

Schultz, T. P. (1984). Studying the impact of household economic and community variables on child mortality. *Population and Development Review 10*, 215–235.

Secretariat, U. N. (2010). *Post Enumeration Surveys Operational Guidelines*. Department of Economic and Social Affairs Statistics Division.

Shen, J. (2013). Increasing internal migration in China from 1985 to 2005: Institutional versus economic drivers. *Habitat International 39*, 1–7.

Shiu, A. and P.-L. Lam (2004). Electricity consumption and economic growth in China. *Energy Policy 32*(1), 47–54.

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society: Series B (Methodological) 31*(3), 539–552.

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University 23*, 20.

Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj 338*.

Sugawara, N., N. Yasui-Furukori, N. Ishii, N. Iwata, and T. Terao (2013). Lithium in tap water and suicide mortality in Japan. *International Journal of Environmental Research and Public Health 10*(11), 6044–6048.

Tetens, J. N. (1786). *Einleitung zur Berechnung der Leibrenten und Anwartschaften: zweyter Theil; Versuche über einige bey Versorgungs-Anstalten erhebliche Puncte.* Bey Weidmanns Erben und Reich.

Theil, H. (1971). *Principles of Econometrics.* New York: John Wiley & Sons.

Tripepi, G., K. J. Jager, F. W. Dekker, and C. Zoccali (2010). Stratification for confounding–part 2: direct and indirect standardization. *Nephron Clinical Practice 116*(4), c322–c325.

Tseng, M.-C. M., I.-C. Cheng, and F.-C. Hu (2011). Standardized mortality ratio of inpatient suicide in a general hospital. *Journal of the Formosan Medical Association 110*(4), 267–269.

Twohig-Bennett, C. and A. Jones (2018). The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. *Environmental Research 166*, 628–637.

Ullah, M., M. Aslam, M. D. M. I. Ullah, and T. LazyData (2019). Package ‘mctest.’.

Wan, G. (2008). *Inequality and Growth in Modern China.* Oxford University Press.

Wang, H. and L. G. Tassinary (2019). Effects of greenspace morphology on mortality at the neighbourhood level: a cross-sectional ecological study. *The Lancet Planetary Health 3*(11), 460–468.

Wang, J. (2013). Trends in life expectancies and moratlity patterns in China since 1990: A further examination and analysis (original language is Chinese). *Population Research 37*(4), 3–18.

Wang, J. and Y. Ge (2013). Assessment of 2010 census data quality and past population changes (original language is Chinese). *Population Research 37*(1), 22–33.

Wang, S. and K. Luo (2018). Life expectancy impacts due to heating energy utilization in China: Distribution, relations, and policy implications. *Science of The Total Environment 610*, 1047–1056.

Wang, S., K. Luo, and Y. Liu (2015). Spatio-temporal distribution of human lifespan in China. *Scientific Reports 5*(1), 1–10.

Wang, W. and L. Lee (2013). Estimation of spatial panel data models with randomly missing data in the dependent variable. *Regional Science and Urban Economics 43*(3), 521–538.

Westergaard, H. (1882). *Die Lehre von der Mortalität und Morbilität: anthropologisch-statistische Untersuchungen*. Fischer.

White, I. R., P. Royston, and A. M. Wood (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine 30*(4), 377–399.

Wilson, D. R. and T. R. Martinez (1997). Improved heterogeneous distance functions. *Journal of artificial intelligence research 6*, 1–34.

Woolf, S. H., R. E. Johnson, G. E. Fryer Jr, G. Rust, and D. Satcher (2004). The health impact of resolving racial disparities: an analysis of US mortality data. *American Journal of Public Health 94*(12), 2078–2081.

Wu, Y., K. Hu, Y. Han, Q. Sheng, and Y. Fang (2020). Spatial characteristics of life expectancy and geographical detection of its influencing factors in China. *International Journal of Environmental Research and Public Health 17*(3), 906.

Wu, Y., Y. Song, and T. Yu (2019). Spatial differences in China's population aging and influencing factors: The perspectives of spatial dependence and spatial heterogeneity. *Sustainability 11*(21), 5959.

Xiang, K. and D. Song (2016). Spatial analysis of China province-level perinatal mortality. *Iranian Journal of Public Health 45*(5), 614.

Xie, Y., H. Dai, Y. Zhang, Y. Wu, T. Hanaoka, and T. Masui (2019). Comparison of health and economic impacts of pm2. 5 and ozone pollution in China. *Environment International 130*, 104881.

Xu, B. and B. Lin (2016). Regional differences of pollution emissions in China: Contributing factors and mitigation strategies. *Journal of Cleaner Production 112*, 1454–1463.

Xu, B. and B. Lin (2018). What cause large regional differences in pm2. 5 pollutions in China? Evidence from quantile regression model. *Journal of Cleaner Production 174*, 447–461.

Yang, M., M. W. Rosenberg, and J. Li (2020). Spatial variability of health inequalities of older people in China and related health factors. *International Journal of Environmental Research and Public Health 17*(5), 1739.

Yang, T.-C., A. J. Noah, and C. Shoff (2015). Exploring geographic variation in US mortality rates using a spatial durbin approach. *Population, Space and Place 21*(1), 18–37.

Yule, G. U. (1934). On some points relating to vital statistics, more especially statistics of occupational mortality. *Journal of the Royal Statistical Society 97*(1), 1–84.

Zhai, H. (2003). Indirect estimation on population mortality based on the 2000 census in China (original language is Chinese). *Population and Economics 2003*(5), 65–69.

Zhang, H., Z. Geng, R. Yin, and W. Zhang (2020). Regional differences and convergence tendency of green development competitiveness in China. *Journal of Cleaner Production 254*, 119922.

Zhang, Y., C. Wang, Y. Wang, Q. Xiao, J. Liu, J. Ma, H. Zhou, J. Pan, Y. Tan, S. Chen, et al. (2018). Mortality from parkinson's disease in China: Findings from a ten-year follow up study in Shanghai. *Parkinsonism & Related Disorders 55*, 75–80.

Zhang, Y.-C., D.-K. Si, and B. Zhao (2020). The convergence of sulphur dioxide (SO2) emissions per capita in China. *Sustainability 12*(5), 1781.

Zhao, B. B. (2012). A modified Lee–Carter model for analysing short-base-period data. *Population Studies 66*(1), 39–52.

Zhao, B. B., X. Liang, W. Zhao, and D. Hou (2013). Modeling of group-specific mortality in China using a modified Lee–Carter model. *Scandinavian Actuarial Journal 2013*(5), 383–402.

Zhao, Z. (2006). Income inequality, unequal health care access, and mortality in China. *Population and Development Review*, 461–483.

Zhou, M., Y. Li, H. Wang, X. Zeng, L. Wang, S. Liu, Y. Liu, and X. Liang (2016). Analysis on life expectancy and healthy life expectancy in China, 1990-2015 (original language is Chinese). *Chinese Journal of Epidemiology 37*(11), 1439–1443.

Zhu, Q., L. Ma, X. Luo, and H. Huang (2012). Toxic epidermal necrolysis: performance of scorten and the score-based comparison of the efficacy of corticosteroid therapy and intravenous immunoglobulin combined therapy in China. *Journal of Burn Care & Research 33*(6), 295–308.

Appendix A

Data

This study is based on 31 provinces situated in mainland China: Beijing (BJ), Tianjin (TJ), Hebei (HB), Shanxi (SX), Inner Mongolia (IM), Liaoning (LN), Jilin (JL), Heilongjiang (HLJ), Shanghai (SH), Jiangsu (JS), Zhejiang (ZJ), Anhui (AH), Fujian (FJ), Jiangxi (JX), Shandong (SD), Henan (HA), Hubei(HB), Hunan (HN), Guangdong (GD), Guangxi (GX), Hainan (HI)[1], Sichuan (SC), Guizhou (GZ), Yunnan (YN), Tibet (XZ), Shaanxi (SN), Gansu (GS), Qinghai (QH), Ningxia (NX) and Xinjiang (XJ). The geographical distribution is shown in Figure A.1.

[1] Hainan is an island and does not have any geographical neighbourhood. We have considered it as a neighbour of Guangxi and Guangzhou, owing to their close economic, cultural and climatic connections.

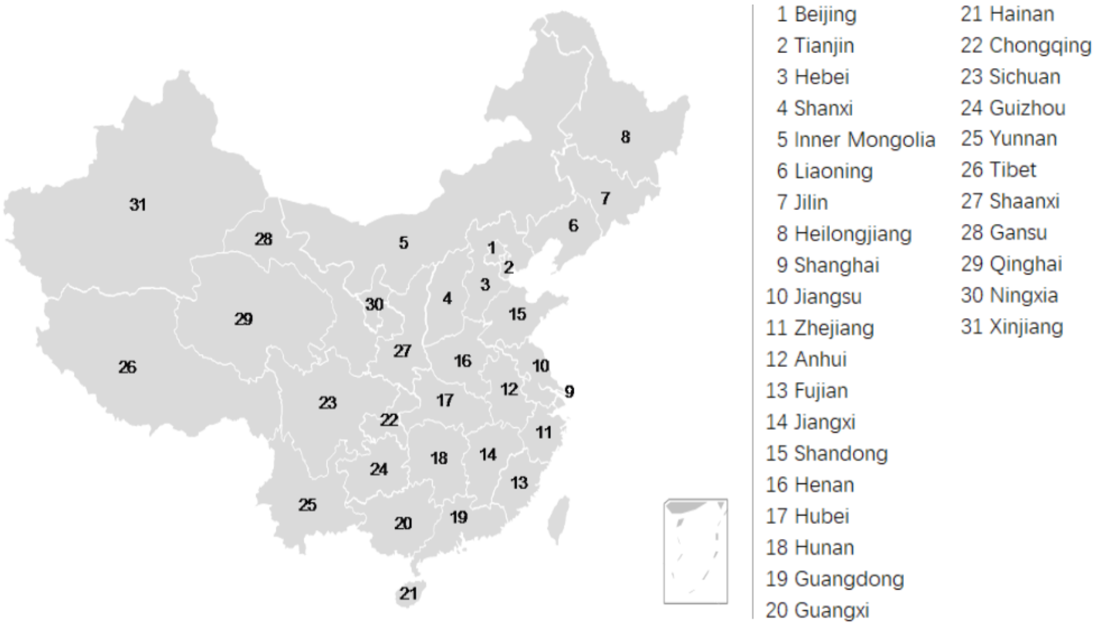


Figure A.1: China's mainland Provincial geography

A.1 Data sources

Data on China was accessed from the following sources:

The National Bureau of Statistics (NBS)[2]

The NBS is responsible for national statistics, national economic accounting and the enactment of statistical regulations. It governs the Provincial Bureaus of Statistics. The NBS provides total populations and crude death rates at the provincial level.

The National Population Census of the People's Republic of China (Census)[3]

A census enumerates every individual in a population, recording information on the national and provincial population and households. Census data collection

[2] http://www.stats.gov.cn/english/

[3] http://www.stats.gov.cn/english/Statisticaldata/CensusData/

is organised by the NBS. Thus far, China has completed six censuses, in 1953, 1964, 1982, 1990, 2000 and 2010, respectively (the 2020 census data is yet to be released). Future censuses will be conducted by the NBS in all years ending with 0. Censuses provide age-wise population and death figures at the provincial level, which were required for this study.

The 1% Population Sample Survey (Micro Census)

Decadal micro censuses in China are also conducted by the NBS. Generally, the NBS conducts a micro census in years ending with 5 (such as 2005 and 2015), and approximately 1% of the population is surveyed. China conducted four micro censuses, in 1987, 1995, 2005 and 2015. The 2005 and 2015 micro censuses provide age-wise population and death figures at the provincial and national levels.

The World Bank[4]

The World Bank (WB) is an international financial institution, and its website provides free global statistical data. It works to improve the quality, timeliness, and relevance of national and international statistics. The WB databases apply international standards and norms and are a consistent, reliable source of information. The WB provides data on total population and population in the age groups 0–14 and 15–64 years for China. It also provides mortality rates for the age groups 0 and 0–5 years and the number of deaths in the age groups 5–9, 10–14, 15–19, and 20–24 years at the national level.

China Population and Employment Statistics Yearbook (PE Yearbook)

It is an informative annual report published by the NBS. With a comprehensive view of China's population and employment situation, it provides the 1‰ population and employment sample survey data for a specific year (except for census and micro census years). It also covers key statistical data for the recent years and some historically important years at the national and provincial level. PE

[4] https://data.worldbank.org/

Yearbooks provide the crude death rates and population at the provincial level. Moreover, they also provide population and death rates by age at the national level.

China Statistical Yearbook (CS Yearbook)[5]

The CS Yearbook is an annual NBS statistical publication, which provides the 1‰ population sample survey results for a particular year (except for census and micro census years). It comprehensively reflects China's economic and social development. CS Yearbooks provide age-wise population at the national level, total population, and crude death rates at the national and provincial levels.

Table A.1 summarises the different data sources for total and age-specific deaths and population at the provincial level. Table A.2 shows the different data sources for total and age-specific deaths and population at the national level. Census and micro census data were used in the study, as data on age-specific deaths and population at the provincial level were available only from these two sources.

[5] http://www.stats.gov.cn/english/Statisticaldata/AnnualData/

	Provincial level			
	Age-specific		Total	
	Number of death/rate	Population	Number of death/rate	Population
NBS	×	×	(rate) 2000-2018	2000-2019
Census	(number) 2000, 2010	1990, 2000, 2010	(number) 2000, 2010	2000, 2010
Micro census	(number) 2005, 2015	2005, 2015	(number) 2005, 2015	2005, 2015
WB	×	×	×	×
PE Yearbook	×	×	(rate) 1990-2019	1990-2018
Yearbook	×	×	(rate) 1999-2019	1999-2018

Note: '×' means that the data is not available for this source.

Table A.1: Data sources for population and deaths at the provincial level

	National level			
	Age-specific		Total	
	Number of death/rate	Population	Number of death/rate	Population
NBS	×	×	(rate) 1949-2019	1949-2019
Census	(both)1990, 2000, 2010	1990, 2000, 2010	(number) 2000, 2010	1990, 2000, 2010
Micro Census	(both) 2005, 2015	2005, 2015	(number) 2005, 2015	2005, 2015
WB	×	×	(rate) 1960-2019	1960-2019
PE Yearbook	1996-2018	1994-2018	(rate) 1949-2018	1949-2019
Yearbook	×	1999-2018	(rate) 1949-2018	1949-2019

Note: '×' means that the data is not available for this source.

Table A.2: Data sources for population and deaths at the national level

A.2 Data description

A.2.1 Mortality Data

The databases for population and number of deaths from censuses in 2000 and 2010 and micro censuses in 2005 and 2015 are summarised as follows:

Year	2000	2005	2010	2015
No. of provinces	31	31	31	31
Population	Census	Micro census	Census	Micro census (1.55%)
Deaths	Census	Micro census	Census	Micro census

Table A.3: Database sources

The age-specific number of death and population were available for 2000, 2005, 2010, and 2015. In this case, the interpolation function was used to construct new data points. The linear interpolations for age-specific deaths and population are expressed as follows:

$$\begin{aligned} n_{i,g}(t) &= n_{i,g}(\tau) + \frac{t-\tau}{5}(n_{i,g}(\tau+5) - n_{i,g}(\tau)), \tau < t < \tau + 5 \\ d_{i,g}(t) &= d_{i,g}(\tau) + \frac{t-\tau}{5}(d_{i,g}(\tau+5) - d_{i,g}(\tau)), \tau < t < \tau + 5 \end{aligned}$$

where year τ may be 2000, 2005, or 2010, $n_{i,g}(t)$ is the population in age group g in province i in year t, $d_{i,g}(t)$ is the number of deaths in age group g in province i in year t.

A.2.2 Variable selection

To explain mortality behaviour, some impact factor variables were considered in the regression model. The data for the 31 Chinese provinces were mainly obtained from the NBS. Table A.4 lists the other data sources:

	Missing NBS data	Solution
Crude birth rate	Heilongjiang (2000) and shaanxi (2000)	Used the 2000 census data
Population growth	Heilongjiang (2000) and shaanxi (2000)	Used the 2000 census data
Dependency ratio	2000, 2001, 2010	Used the 2000 and 2010 census data and performed linear interpolation to find the missing values for 2001 values in 2001
Public green area	Beijing (2005) Tianjin (2007, 2008)	Used linear interpolation to find the missing values
SO_2 emissions	2000-2003	Used Yearbook data 2001–2004
GRP per capita	No	-
Household consumption level	No	-
Urbanisation	2000-2004	Used the 2000 census data Used linear interpolation to find missing values from 2001 to 2004
Electrical consumption	Tibet (2000-2005)	Assuming the electricity consumption in Tibet increased by around one unit per year from 2000 to 2009, the missing values from 2000 to 2005 were then interpolated.

Table A.4: Missing data in NBS and its solutions

The impact factor variables depend on four aspects: demography, environment, economy, and societal development.

- **Demography** - China was recognised as an ageing society from 2000 when the proportion of the population above 65 years of age exceeded 7%. The domestic migration from less-developed provinces to more-developed provinces changed the demographic distribution in different provinces (Shen, 2013). Over the last four decades, many people in China, including millions of parents from small cities or rural areas, have migrated to larger cities for jobs, leaving their children (termed 'left-behind children') with grandparents. This phenomenon further diversifies the age-specific population distribution between provinces as well as urban and rural areas. Hence the crude birth rate and population growth are different at the provincial level of China. Meanwhile, it makes dependency ratios to be different at the provincial level of China. Crude birth rate, population growth and dependency ratio are considered as impact factor variables for mortality in the study.

- *Crude birth rate (‰)*: It is the total number of live births per 1,000 population for a reference period.

- *Population growth (‰)*: It is the increase in the number of individuals in a population. Specifically, it refers to the change in population over unit time.

- *Dependency ratio (%)*: It is the age-wise population ratio of those typically not in the labour force (0–14 and 65+ years) to those typically in the labour force (15–64 years). It is used to measure the pressure on the productive population.

- **Environment** - The green area and air pollution are relevant to human health. Wang and Tassinary (2019) pointed out that green spaces in cities and mortality risk have an obvious correlation and the level of the green area has large differences among provinces in China (Zhang et al., 2020). Hence, the green spaces impact mortality in the different level at the provincial level of China. Besides, SO_2 has several negative effects on human health (Lin et al., 2004; Khan and Siddiqui, 2014; Goudarzi et al., 2016). China is the third-largest SO_2 emitting country in the world. Moreover, The levels of SO_2 show geographical diversity in China and it is associated with increased premature mortality and morbidity (Cao et al., 2012). Hence, public green area and SO_2 emission are considered as impact factor variables for mortality in the study.

- *Public green area (m^2 per capita)*: The per capita public green area is an important indicator reflecting the living environment and quality of life of the residents and which includes district and residential parks and botanical gardens.

- *Sulphur dioxide (SO_2) emissions (kg per capita)*: It includes both industrial and residential emissions. The former mainly result from the burning of fossil fuels by power plants and other industrial facilities. The second latter result from domestic consumption and human activities, such as heating and kitchen ranges.

- **Economy** - The development of the economy impacts mortality rates and its impacts on poor and rich regions are very different. The life expectancy of poor areas is easier influenced by incomes than in rich areas (Linden and Ray, 2017). The increase of incomes will improve life quality in poor areas. Meanwhile, the high income areas are easier to provide the infrastructure for public health in order to maintain people's health (Cutler et al., 2006). In China, the mortality rates are disparity between more developed provinces and less developed provinces (Zhao, 2006). Hence, GDP per capita and household consumption level as indexes to measure people's life quality among countries (or areas, provinces) are considered as impact factor variables for mortality in the study.

- *Gross regional product per capita (1,000 yuan per capita)*: GRP is simply the GDP for each province. It is a monetary measure of the market value of all final goods and services produced by a particular province.

- *Household consumption level (1000 yuan)*: Household consumption is the total consumption expenditure by households on their everyday needs, such as food, clothing, housing, energy, transport, and health costs.

- **Societal development** - Urbanization brings high quality of residential environment, education resources and medical treatments, which indirectly impacts mortality (Muller, 2002; Lleras-Muney, 2005). Similarly, urbanization also influences mortality rates in China through other aspects (Luo et al., 2015; Li et al., 2016; Hou et al., 2019). However, urbanization levels at the provincial level of China are different (Lin et al., 2018), which means mortality rates may have been influenced differently by urbanization level at the provincial level of China. Besides, electricity consumption has a closed relationship with economic growth in China (Shiu and Lam, 2004). Mazur (2011) mentioned that the increase of electricity consumption is essential for poor areas and countries to improve well-being. In China, all rural areas have been electrified since 2018[6]. Not surprisingly, the electricity consumption is unbalanced at the provincial level in China, in the sense that provinces with more rural areas particularly have less electricity consumption. Hence, ur-

[6] Report from NBS: http://www.stats.gov.cn

banization and electrical consumption per capita are considered as impact factor variables for mortality.

- *Urbanisation (%)*: It is defined as the proportion of a population living in urban areas.

- *Electrical consumption (1,000 kwh per capita)*: Electrical consumption is included in industrial and residential consumption. In the present study, electrical consumption impacts the economy and daily life.

A.3 Missing population census data

Population censuses are aimed at deciphering a country's demographic structure. However, worldwide, errors and inconsistencies invariably occur during large population censuses. Generally, errors made during population censuses include omissions, duplications, and erroneous inclusions (Secretariat, 2010). The first type refers to missing housing units, households, or persons in census enumeration. The second happens when housing units, households, or persons are recorded multiple times. The last occurs if housing units, households, or persons are enumerated but in incorrect locations. Gross coverage error is a combination of omissions, duplications, and erroneous inclusions.

China's population enumeration is a challenging task as it has the largest population. Omission of births and deaths is one of main problems in population censuses in China. First, families may not report new-borns to escape the punishment for violating the one-child policy. They may also not report the deaths of their family members because they would like to retain pension benefits. Second, death, especially a new-born's, is believed to be ominous in Chinese culture, which makes families reluctant to announce such deaths. Finally, the enumeration of deaths uses the permanent residence rather than the household registration ('Hukou' in Chinese) location. For example, when a person living alone in Beijing, with his hukou in Shanghai, passes away, the death will be counted in Beijing's data, and not Shanghai's. However, as the deceased lived alone, this case may not be reported to the Beijing government. Duplications occur when persons are enumerated more than once. A person may have been counted in the province of their permanent residence in the 2000 census, but simultaneously in their current living location and hukou location in the 2010 census.

The evaluation of census data in China focuses on the population distribution and the number of deaths. To evaluate the 2000 census, Zhai (2003) calculated the omission of death rates in the 2000 population census. For the 2010 census, Wang (2013) applied the cohort survival and brass logic life table methods to re-estimate mortality rates, not only for 0–4 year olds but also for those over 60 years of age. Wang and Ge (2013) introduced the forward survival method to evaluate omission and duplication of population by gender and age in the 2010 census. Hongyang Cui (2013) used demographic analysis to assess the 2010 census based on the 2000 census. Chen and Zhang (2015) adjusted the population with aged 0-5 years in the 2010 population census and then evaluated the fertility rate.

Age-specific deaths across sexes for different provinces in 2005 and 2015 are provided in the 2005 and 2015 micro censuses, which contain data for 31 provinces. However, some values were found missing in the data set. There may be two kinds of missing data in one age group in a province. The first case is that the data for only one sex (male or female) is missing, and the second is that the data for both males and females is missing. Notably, the number of deaths of males and females in the age group 1–4 was not available for Tianjin (TJ) in 2005, unlike the total death numbers for this age group. This may either be an error, or it may be the case that the gender information was missing and only the sum was recorded. Hence, we classify it as the second type of missing data. For the summary of missing data, see tables A.7, A.8 ,A.5 and A.6; the shaded parts indicate that the data for the age group are missing for the respective province.

		BJ	TJ	HB	IM	XZ	QH	NX
100+	T							
	M	▨	▨	▨		▨	▨	▨
	F				▨			

Table A.5: Missing data of population in 2005

		TJ	XZ	QH	NX
100+	T				
	M	▨	▨	▨	▨
	F				

Table A.6: Missing data of population in 2015

		BJ	TJ	SX	IM	JL	HLJ	SH	JS	ZJ	HA	GX	HI	CQ	XZ	SN	GS	QH	NX
1-4	T	▨																	
	M	▨	▨																
	F	▨	▨			▨													
5-9	T							▨											
	M	▨						▨											
	F		▨		▨			▨											
10-14	T																		
	M	▨																	
	F							▨					▨						
15-19	T																		
	M																		
	F							▨											
20-24	T																		
	M							▨											
	F	▨																	
95-99	T														▨			▨	
	M				▨		▨								▨			▨	
	F														▨			▨	▨
100+	T			▨											▨			▨	▨
	M	▨		▨			▨	▨	▨	▨	▨	▨	▨	▨	▨		▨	▨	▨
	F		▨	▨	▨										▨	▨		▨	▨

Table A.7: Missing data of death number in 2005

		BJ	TJ	IM	LN	JL	HLJ	SH	JS	ZJ	AH	FJ	SD	HB	HI	GZ	XZ	SN	GS	QH	NX	XJ
0	T																					
	M				▨																	
	F	▨	▨			▨																
1-4	T																					
	M						▨															
	F	▨				▨		▨	▨					▨								
5-9	T	▨						▨														
	M	▨	▨					▨	▨													
	F	▨	▨			▨	▨	▨				▨								▨	▨	
10-14	T	▨	▨			▨	▨															
	M	▨	▨			▨	▨															
	F	▨	▨	▨	▨	▨	▨	▨		▨	▨	▨		▨	▨						▨	
15-19	T	▨						▨														
	M	▨	▨					▨														
	F	▨						▨	▨	▨											▨	
20-24	T																					
	M			▨																		
	F	▨				▨															▨	
30-34	T																					
	M																					
	F	▨			▨																▨	
95-99	T																▨					
	M																▨					
	F																▨					
100+	T																▨			▨		
	M			▨									▨	▨	▨	▨	▨				▨	
	F						▨										▨	▨	▨	▨		▨

Table A.8: Missing data of death number in 2015

A.4 Data imputation

Data collected from a large-scale survey often has missing values. For example, some individuals are not willing to respond questions about personal life, especially in death information. Missing values have a significant impact on descriptive statistics and predictive analysis. Micro censuses in China also face challenges from missing data. According to Rubin (1976) classification, missing data types are classified as missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). We denote U is our data set, where $U = \{U_{\text{obs}}, U_{\text{mis}}\}$, U_{obs} and U_{mis} are the observed and missing data in U, respectively. The MCAR indicates that the missing data are independent of the observed and unobserved data. Hence we have

$$f(M|U,\phi) = f(M,\phi)$$

where M is the missing data and ϕ is the unknown parameter which can completely characterises the data U.

Similarly, the MAR shows that the missing data is related to the observed but not unobserved data, it denotes as

$$f(M|U,\phi) = f(M|U_{\text{obs}},\phi)$$

NMAR means the missing data depends only on the missing data themselves, and not on any other observed data.

$$f(M|U,\phi) = f(M|U_{\text{mis}},\phi),$$

A.4.1 Imputation methods in general

The simplest way to handle missing values is to delete the missing value in order to obtain complete data. However, it discards a lot of information in these missing data, which may lead to biased or wrong conclusions. There are quite a few imputation algorithms proposed to solve missing data problems in the literature. The most widely used ones can be found in the following.

- **Mean imputation**

  It replaces missing values by the mean of the available observations. This method keeps the data size. However, the variability in the data and the standard deviations and the variance estimates may be underestimated.

- **Regression imputation**

  It fills missing values with predicted values from a regression model by using the information in the complete data. This method assumes that the imputed values rely on a regression line. As a result of imposing linearity, this method will overestimate the correlation between independent and dependent variables, and at the same time, their variances and covariances will be underestimated (Enders, 2010).

- **Multiple imputation**

  This method is a simulation-based procedure, which imputes multiple data from the distribution of the observed data. There are mainly three steps to implementing this method. Firstly, one needs to create multiple datasets with the missing values being replaced by imputed values. In other words, the missing values are filled in multiple times to create multiple complete datasets. Secondly, the multiple complete datasets are analysed by standard statistical methods to fit the model. Subsequently, the parameter estimates from each imputed value are combined to obtain final results (Sterne et al., 2009). Multiple imputation is essentially an iterative form of stochastic imputation. Each imputed data includes a random component whose magnitude reflects the extent to which other variables in the imputation model cannot predict its true value (Johnson and Young (2011),White et al. (2011)).

- **Hot-deck method**
  The missing data is replaced with another value from other complete observations. The data set with missing values is called the recipient and the one with complete observations is called the donor. This method is constituted by two steps. Firstly, the data set is divided into several clusters and each component with a missing value is associated with one cluster. The complete component in this cluster is applied in filling in the missing data. This method does not depend on model fitting and thus is potentially less sensitive to model misspecification (Andridge and Little, 2010). In addition, when the donor in the hot-deck method is selected randomly from a set of potential donors (called donor pool), this method is referred to as the random hot-deck method (Andridge and Little, 2010).

The variance of a mean-imputed variable is always biased downward from the variance of the un-imputed variable. Mean imputation reduces the variance of the imputed variables, reduces standard errors, which invalidates most hypothesis tests and the calculation of confidence interval and does not preserve relationships between variables (Enders, 2010). On the other hand, the validity of the multiple imputation results is questionable if there is an incompatibility between the imputation model and the underlying model, or if the imputation model is less general than the underlying model. This means that the imputed values in multiple imputation depends on how appropriate modelling is done, which makes it challenging to successfully implement it in practice (Soley-Bori, 2013). Hence the hot-deck method and regression imputation are considered in the study and they will be discussed in detail in the following.

A.4.2 Considered imputation methods

In this study, we applied the k nearest neighbours imputation, the linear trend of age-specific death rate and gender ratio to estimate missing data.

A.4.2.1 k Nearest neighbours (kNN) imputation

Nearest neighbours (NN) imputation is a donor-based method in which the imputed value is either a data point that was actually measured from another record in a database (1-NN) or the average of measured data from k records (kNN) (Beretta and Santaniello, 2016). The kNN imputation is a kind of hot-deck method with k donors being selected from the neighbours. The most notable characteristics of NN imputation are the followings. Firstly, the imputed values are occurring values and not constructed values. Secondly, this method makes use of auxiliary information provided by the independent variables, thus preserving the original structure of the data. In addition, it is fully non-parametric and does not require explicit models to relate independent and dependent variables, thus being less prone to model misspecification. The process of applying the kNN method is as follows.

- We divide each province with missing values, it is divided into two parts A_{obs} and A_{mis}

  $$A = A_{obs} + A_{mis}$$

  where A_{obs} is age-specific with observed data, A_{mis} is age-specific groups with missing values and A is the total age-specific groups.

- For each missing data point, the k nearest neighbours are chosen as donors where the distance is in the Euclidean sense. Then, we use the median/mean of these k nearest neighbours to impute the missing value, i.e.

  $$A_{knn} = A_{obs} + A_{fill}$$

where A_{fill} is age-specific groups of imputed values and A_{knn} is the total age-specific groups without missing values.

It is important to find the value of k in kNN imputation. Duda et al. (1973) mentioned that the optimum k is chosen to be $\sqrt{N}$, where N in our case corresponds to the number of provinces. Some other authors suggested that k should be low (1 or 2). The imputation would be too sensitive when k is set to be 1 since the replaced value is fully determined by the nearest neighbour. So, Cartwright et al. (2004) suggested that k should be 2. In this study the idea of $\sqrt{N}$ is chosen and the resulting value of k is $5 \approx \sqrt{31}$.

To determine k nearest neighbours, a distance function is applied. The most common choice is Euclidean distance function (Wilson and Martinez, 1997). Besides, the Gower distance (Kowarik and Templ, 2016) and Mahalanobis distance functions are other options in kNN imputation (Maltamo et al., 2003). In this study, the Euclidean distance function can be defined as

$$d(a,b) = \sqrt{\sum_{g \in G} (x_{ag} - x_{bg})^2}$$

where $d(a,b)$ is the distance between province a and b, x_{ag} and x_{bg} are the death numbers at g age group in province a and b, G is the set of age group with non-missing values in both cases. In general, the kNN algorithm has several advantages. It is simple to implement, intuitive to understand, and does not require any training time compared with other more advanced algorithms. Because of this, as we keep adding new data to the dataset, the prediction is adjusted without having to retrain a new model. The choice of distance functions also brings some flexibility since apparently different values will be used when the nearest neighbours are no longer the same ones. In addition, the method can predict both discrete and continuous data. Meanwhile, it does not create a predictive model for each component with a missing value.

A.4.2.2 Regression imputation

We assume that both the proportion of death number in each age-specific group and the gender ratio across years have the linear trend.

Data interpolation for 2005

When only death (or population) numbers for one sex (male or female) are unknown, it is assumed that the age-specific death (or population) ratio for this sex in 2000, 2005, and 2010 has a linear trend. Hence, the missing male/female data for 2005 for age group g in province i can be evaluated as follows:

$$\frac{Q_{i,g}(2000) + Q_{i,g}(2010)}{2} = \frac{M_{i,g}(2005)}{F_{i,g}(2005)}, \tag{A.1}$$

where $Q_{i,g}(2000)$ and $Q_{i,g}(2010)$ are the male to female ratios of death (or population) numbers in age group g for province i in the years 2000 and 2010, respectively. $M_{i,g}(2005)$ and $F_{i,g}(2005)$ are the death (or population) numbers for males and females in age group g in province i in 2005. For example, the male to female ratio of deaths in the second age group for Beijing (BJ) in years 2000 and 2010 were 1.3 and 1.1, respectively. The number of male deaths in the second age group in Beijing in 2005 was 300, and the number for female deaths was unknown. We then have $\frac{1.3+1.1}{2} = \frac{300}{F_{i,g}(2005)}$ and $F_{i,g}(2005) = 250$.

Once the missing data is calculated, the total number of deaths in these age groups must be adjusted according to the 'new' data. The next step involves the case that both genders have missing values, as the following method uses the sum of all known deaths (including the 'new' data obtained using Equation A.1) in a province.

The following method is used in the case wherein both female and male data values are missing. We assume that the proportion of deaths in each age group has a linear trend. Suppose there are a total of G age groups and m $(0 \leq g \leq G)$ age groups have unknown number of deaths in a province i in year t , denoted as $x_{i,1}(t), \cdots, x_{i,g}(t)$. $A_{i,G-m}(t)$ is the sum of deaths apart from $x_{i,1}(t), \cdots, x_{i,g}(t)$ in province i in year t, where

$$\sum_{k=1}^{g} x_{i,k}(t) + A_{i,G-g}(t) = \sum_{k=1}^{G} x_{i,k}(t).$$

$B_{i,k}(t)$ is the proportion of deaths for k age group for province i in year t.

$$\begin{aligned} B_{i,k}(t) &= \frac{x_{i,k}(t)}{\text{Total values for all age-specific groups}} \\ &= \frac{x_{i,k}(t)}{\sum\limits_{k=1}^{g} x_{i,k}(t) + A_{i,G-g}(t)} \end{aligned}$$

where $k = 1, \cdots, g$.

We estimate these unknown deaths in 2005 by solving for the following system of linear equations. For age group k, we have

$$\frac{B_{i,k}(2000) + B_{i,k}(2010)}{2} = \frac{x_{i,k}(2005)}{A_{i,G-g}(2005) + \sum\limits_{k=1}^{g} x_{i,k}(2005))},$$

where $k = 1, \ldots, g$.

Data interpolation for 2015

The 2005 interpolation method does not apply to 2015 because 2020 data are not available currently. The life expectancy in China has continuously increased in recent decades. Hence, for provinces that have missing values of age-specific deaths, we use the smallest value of the age-specific death rate g for province i in 1990, 2000, and 2010 to approximate the age-specific missing values in 2015. Therefore, age-specific death numbers can be obtained by the following equation,

$$d_{i,g,s}(2015) = n_{i,g,s}(2015) \cdot \min\left\{r_{i,g}(1990), r_{i,g}(2000), r_{i,g}(2010)\right\}$$

where $s = \{\text{female, male, female and male}\}$. $d_{i,g}$ is the g age-specific death number in province i, $n_{i,g}$ is the g age-specific population in province i and $r_{i,g}$ is the g age-specific death rate in province i.

A.4.3 Imputation evaluation

It is necessary to measure how the choice of imputation methods would affect our analysis results and hence to select the most appropriate imputation in this study. After dealing with missing data in micro censuses, mortality standardised and spatial panel data model will be used in further analysis.

The SMR and CMF for all provinces are calculated with original data (missing values are filled in zero), kNN (from section A.4.2.1) and regression imputation (from section A.4.2.2). Table A.9 and A.10 summarise minima, the first quartiles, medians, means, the third quartiles and maxima of SMR and CMF in 2005 and 2015 respectively. At the same time, it shows how values vary with different imputed methods by calculating the variation ratio $\frac{max-min}{min}$. These ratios are between 0 and 17.62% in SMR and are between 0.02% and 18.44% in CMF. After this the boxplot is used to find which province is influenced the most by the choice of imputation method. The boxplot is a standardised way of displaying the distribution of data. It shows five basic statistics: minimum, first (or lower) quartile (Q1), median, third (or upper) quartile (Q3), and maximum (details see Figure A.2). The interquartile range is the length of the middle 50% of the interval of space. The first and third quartiles are 25% and 75% of the data points respectively. The data point that is located outside the whiskers of the box plot is called an outlier. In Figure A.3, the red boxplot shows the values of SMR when the kNN imputation is applied, the blue boxplot shows the values of SMR when the regression trend is assumed, and the green boxplot shows the values of SMRs with original data,

	SMR_{linear}	SMR_{knn}	$\text{SMR}_{original}$	rate
Minimum	0.640	0.652	0.637	0.0234
First quartile	0.926	0.926	0.925	0.0006
Median	0.960	0.960	0.960	0.0000
Mean	1.029	1.034	1.028	0.0064
Third quartile	1.136	1.136	1.135	0.0010
Maximum	1.662	1.733	1.657	0.0456
	CMF_{linear}	CMF_{knn}	$\text{CMF}_{original}$	rate
Minimum	0.605	0.636	0.600	0.0600
First quartile	0.927	0.928	0.926	0.0012
Median	0.968	0.968	0.968	0.0002
Mean	1.017	1.026	1.016	0.0107
Third quartile	1.140	1.142	1.137	0.0044
Maximum	1.488	1.613	1.479	0.0906

Table A.9: SMR and CMF in 2005 with data using different imputation methods

	SMR_{linear}	SMR_{knn}	$\text{SMR}_{original}$	rate
Minimum	0.648	0.670	0.640	0.0464
First quartile	0.892	0.893	0.889	0.0040
Median	1.021	1.028	1.023	0.0066
Mean	1.045	1.056	1.041	0.0152
Third quartile	1.169	1.169	1.169	0.0000
Maximum	1.630	1.900	1.615	0.1762
	CMF_{linear}	CMF_{knn}	$\text{CMF}_{original}$	rate
Minimum	0.640	0.677	0.630	0.0758
First quartile	0.889	0.891	0.887	0.0051
Median	1.016	1.018	1.015	0.0025
Mean	1.032	1.050	1.028	0.0205
Third quartile	1.171	1.184	1.171	0.0109
Maximum	1.552	1.838	1.552	0.1844

Table A.10: SMR and CMF in 2015 with data using different imputation methods

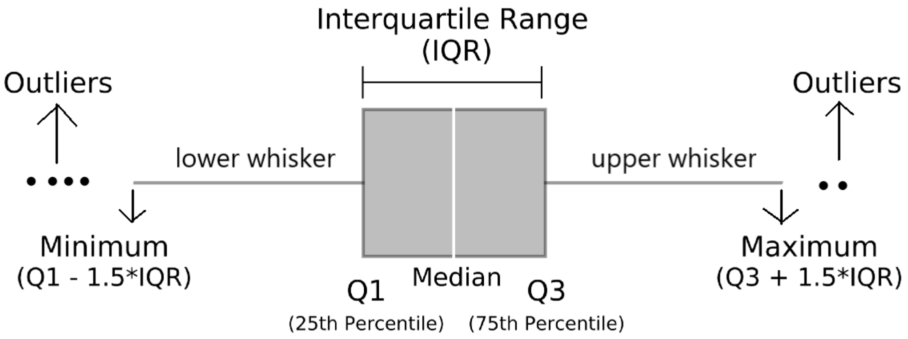


Figure A.2: The boxplot structure

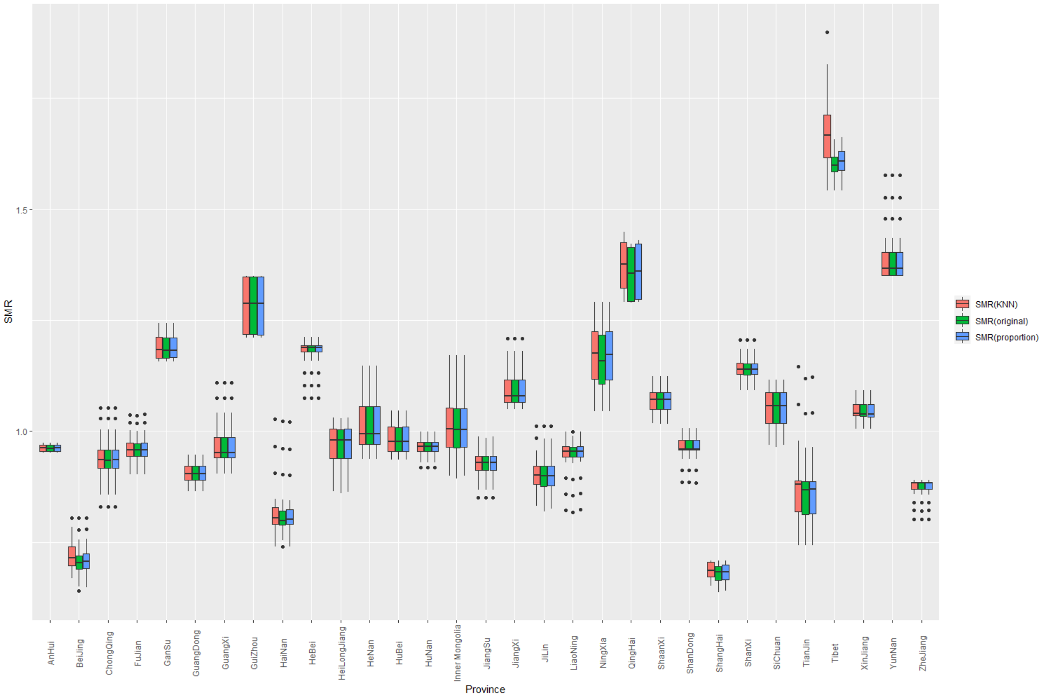


Figure A.3: SMRs with different miss data imputation method

which means the missing values are simply replaced by zeros. Figure A.4 shows a similar boxplot when using CMF instead of SMR. It is found that values of SMR and CMF are not influenced by the imputation method of missing values in almost all provinces, except Tibet.

Here is one potential reason why Tibet is the special one. Table A.11 shows SMR,

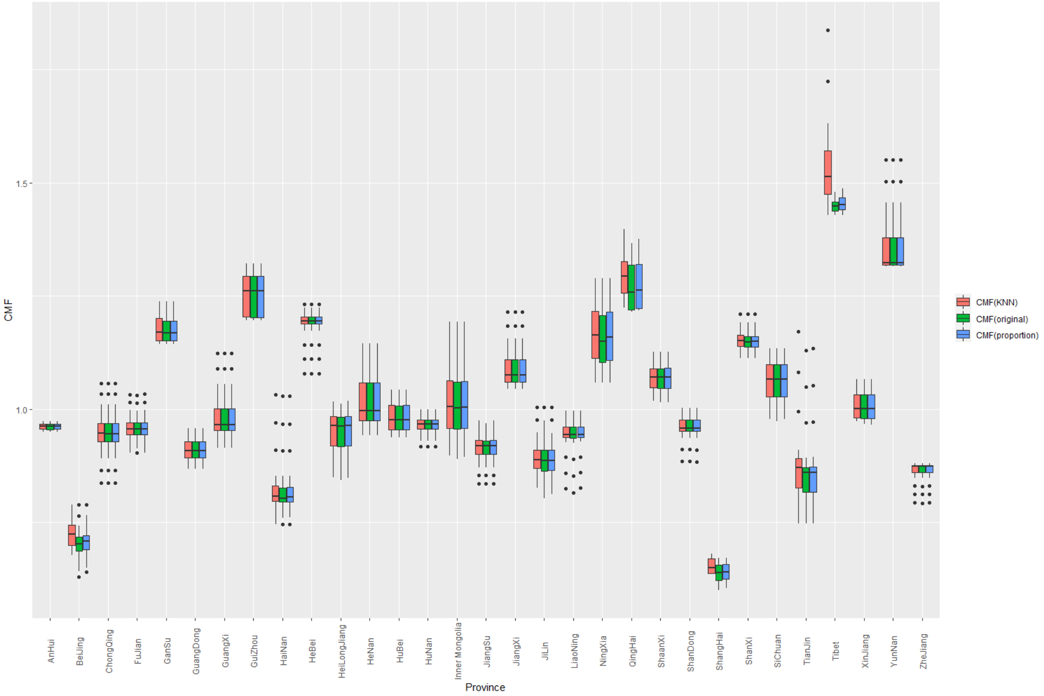


Figure A.4: CMFs with different miss data imputation method

CMF and imputed death numbers in Tibet in age groups 20-24, 30-34, 95-99 and above 100 in the years 2005 and 2015 using *k*NN imputation and regression assumption. It can be found that imputed values in age groups 20-24 in 2005 and age group 20-24 in 2015 are very similar no matter what kinds of methods are used in imputation. However, imputed values are totally different at ages above 95 years in both 2005 and 2015. The imputed death numbers from *k*NN imputation are much larger than deaths under the regression assumption. In specific, the imputed death number at age 95-99 is 2129 with *k*NN imputation, but is 128 under the regression assumption. It suggests that the *k*NN imputation at age above 95 in Tibet may not be reasonable since the life expectancy at birth in Tibet is 68 years in 2010 and 2029 is a large death number. In the next step, we compare death numbers above age 95 in 2000 and 2010 with imputed death values above age 95 in 2005 and 2015 in Tibet, since death numbers in 2000 and 2010 are available.

	Imputed method	20-24	30-34	95-99	100+	SMR	CMF
2005	Deaths$_{knn}$	301	-	679	226	1.733	1.613
	Deaths$_{linear}$	302	-	45	13	1.662	1.488
2015	Deaths$_{knn}$	-	516	2129	452	1.900	1.838
	Deaths$_{linear}$	-	516	128	10	1.630	1.552

Table A.11: Missing values imputation in Tibet

Table A.12 shows the death number at ages above 95 years from 2000 to 2015 in Tibet. The death number at age above 95 years in 2000 and 2010 from Census 2000 and 2010, the death number in 2005 and 2015 are imputed by regression assumption and kNN method. It is found that the imputed values under the linear assumption are reasonable compared with imputed values from the kNN method. Hence, the assumption of linear trend is used to handle missing data in the study.

Age group	Death number	2000	2005	2010	2015
95-99	Deaths$_{kNN}$	-	679	-	2129
	Deaths$_{linear}$	-	45	-	128
	Deaths$_{original}$	36	-	40	-
100+	Deaths$_{kNN}$	-	226	-	452
	Deaths$_{linear}$	-	13	-	10
	Deaths$_{original}$	10	-	12	-

Table A.12: Deaths with ages above 95 in Tibet from 2000 to 2015

Similarly, imputation methods will not influence the spatial panel data model. After using VIF to avoid multicollinearity, it is found that different imputation algorithms do not change the result of VIF. Then same independent variables with different dependent variables will be used in the spatial panel data model to find which model fits the data best. Table A.13 shows the result of model selection. The dependent variables are logarithms of SMR and CMF. Three sets of data are used: the original one after filling missing values in zero, the complete ones using kNN imputation and linear assumption respectively. The model is selected when the null hypothesis was rejected at the 10% level of the significance level. The table shows that the spatial lag model with time-specific effect should be used in the study, since the p-value is less than 0.10. It is found that log(SMR) is chosen as the independent variable no matter which data set is considered. Hence, the imputation methods for missing data do not really change the result in the spatial panel data model.

	Dependent variable	Test	Space-specific effect μ_i		Time-specific effects ν_t		Spatial-specific and time specific effects	
			Value	p-value	Value	p-value	Value	p-value
Linear	log(SMR)	LM(lag)	0.229	0.633	3.275	0.070	0.124	0.725
		LM(error)	0.241	0.624	0.673	0.412	0.049	0.826
	log(CMF)	LM(lag)	0.193	0.661	1.158	0.282	0.139	0.709
		LM(error)	0.141	0.708	1.337	0.248	0.045	0.832
kNN	log(SMR)	LM(lag)	0.547	0.460	3.582	0.058	0.279	0.597
		LM(error)	0.384	0.536	0.528	0.468	0.145	0.703
	log(CMF)	LM(lag)	0.660	0.416	1.600	0.206	0.295	0.587
		LM(error)	0.310	0.578	0.918	0.338	0.152	0.697
Original	log(SMR)	LM(lag)	0.322	0.570	3.529	0.060	0.243	0.622
		LM(error)	0.397	0.529	0.534	0.465	0.176	0.675
	log(CMF)	LM(lag)	0.259	0.610	1.345	0.246	0.225	0.635
		LM(error)	0.267	0.605	1.101	0.294	0.137	0.711

Table A.13: Section of the spatial panel data model