PLOS ONE

RESEARCH ARTICLE

Semiparametric change points detection using single index spatial random effects model in environmental epidemiology study

Hamdy F. F. Mahmoud[1,2]*, **Inyoung Kim**[1]

1 Department of Statistics, Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, VA, United States of America, **2** Department of Statistics, Mathematics, and Insurance, Assiut University, Asyut, Egypt

* ehamdy@vt.edu

OPEN ACCESS

Citation: Mahmoud HFF, Kim I (2024) Semiparametric change points detection using single index spatial random effects model in environmental epidemiology study. PLoS ONE 19(12): e0315413. https://doi.org/10.1371/journal.pone.0315413

Editor: Laleh Tafakori, RMIT University, AUSTRALIA

Received: February 22, 2024

Accepted: November 25, 2024

Published: December 12, 2024

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Environmental health studies are of great interest in research to evaluate the mortality-temperature relationship by adjusting spatially correlated random effects as well as identifying significant change points in temperature. However, this relationship is often not expressed using parametric models, which makes identifying change points an even more challenging problem. This paper proposes a unified semiparametric approach to simultaneously identify the nonlinear mortality-temperature relationship and detect spatially-dependent change points. A unified method is proposed for the model estimation, spatially dependent change points detection, and testing whether they are significant simultaneously by a permutation-based test. We operate under the assumption that change points remain constant, yet acknowledge the uncertainty regarding their precise number. These change points are influenced by the smoothing of an unknown function, which in turn relies on a smoothing variable and spatial random effects. Consequently, the detection of change points may be influenced by spatial effects. In this paper, several simulation studies are conducted to evaluate the performance of our proposed approach. The advantages of this unified approach are demonstrated using epidemiological data on mortality and temperature.

Introduction

For centuries, the effects of weather and global warming on people have been a public health concern. Previous studies [1–3] have indicated that the temperature-mortality relationship can be depicted as a U, J, or V curve; that is, episodes of extremely hot or cold temperatures increase mortality. The lowest end of the curve was defined as the minimum mortality temperature or the change point—that is, the temperature of the lowest mortality. Extreme temperatures increase the heart rate because of the increase of blood flow from the body to the skin, which can lead to shaking in cold temperatures or sweating in high temperatures. The human body has multiple thermoregulatory mechanisms to counter extreme heat and cold conditions to keep temperature homeostasis within normal values. When temperature change occurs within certain ranges, the human body can adapt and allow individuals to follow some physical

and mental activities, but exposure to temperature extremes outside these ranges for a long period is a risk to human health and can result in mortality. According to [4], elevated mortality rates correlate with high temperatures, primarily attributed to illnesses such as cerebrovascular, cardiovascular, and respiratory diseases. This phenomenon is attributed to the effect of hot temperatures on raising blood cholesterol and viscosity levels.

Climate change is a serious public health issue, and specific policies to reduce the effects of heat waves would be appropriate for public policy. These policies need to target successful interventions and populations that are vulnerable. One of the possible mitigation strategies for this is using air conditioning. Because climate change will likely increase the mean temperature, as well as the frequency of heat events, it is very important to evaluate the links between human health and climate, to better identify populations at risk and take preventive measures. As mean temperatures continue to rise in the future, the issue of heat-related mortality is poised to escalate. By delving into the connection between temperature and mortality rates, as well as identifying change points within cities, we can enhance awareness surrounding hot weather as a significant environmental hazard.

One city

Many articles have studied the mortality-temperature relationship in a specific area or city [5–10]. In these studies, the nonlinear mortality function is first estimated by the generalized linear model and then the change point is detected by observing the temperature degree that is associated with the minimum risk. No testing of whether the change point is statistically significant is considered. [11] studied the mortality-temperature function in a single city, Seoul City, South Korea, using the single index model and tested the significance of the change point by a permutation-based test.

Multiple cities

Some articles [12, 13] have studied multiple cities and have found that change points were associated with temperature and they varied by location, especially with latitude, people who live in cities at higher latitudes have lower thresholds for ambient temperature, whereas people who live in cities at lower latitudes have higher thresholds for ambient temperature [4, 14–16]. In these studies, the generalized additive model is used to estimate the temperature-mortality relationship for each city separately, and the minimum mortality risk or AIC criterion is used to find the change point. Other studies have used a distributed lag nonlinear model to estimate the relationship in each city. [17–20] studied 15 European cities, 63 cities in five East-Asian Countries, 47 Japanese cities, and 31 Chinese cities, respectively. After estimating the relationship in each city separately, the change point is estimated as the temperature that is associated with minimum mortality or maximum likelihood.

However, these studies have not fully addressed (1) whether the change points are accurately detected and tested in multiple cities cases, (2) whether spatial random effect plays an important role in the model, and (3) whether the model assumption is flexible in terms of the link function compared to the single index model when multiple cities considered.

Problem and objectives

This paper introduces a unified semiparametric approach to simultaneously identify the nonlinear mortality-temperature relationship, and detect and test spatially-dependent change points. This approach includes a proposed model and a permutation test. To the best of our knowledge, no such model has been introduced in the statistical literature. We refer to this

model as the "semiparametric change points single index spatial random effects model" (CP-SISM). The proposed approach has the following four characteristics:

1. Spatial random effects are incorporated into the model not only because ignoring random effects may mask the true form of the mortality-temperature relationship due to aggregating the data of all cities, but also to make the proposed model able to predict mortality at new locations. The six cities in our motivating data are located close to each other due to the size of South Korea, so we assume that the correlation between spatial effects exists. Hence, the detection of change points can be affected by spatial effects. Previous work studied each city separately without including the spatial effect.
2. The model is flexible in terms of the link between the response variable and the mean function. A semiparametric approach, based on the single index model, is employed to simultaneously estimate the nonlinear mortality-temperature relationship while adjusting for weather variables. The single index model is chosen because it combines parametric and nonparametric components, offering a flexible representation of real data and enabling the proposed model to effectively describe nonlinear relationships. This approach also helps avoid misleading results that can arise from selecting an incorrect link function. Previous studies often utilized generalized linear models or additive models to estimate the temperature-mortality relationship.
3. The change points are included in the nonparametric function to ensure accurate detection. In the proposed model, change-point parameters are incorporated into the single index function because smoothing the unknown mean function may impact change-point detection. In previous studies, the change point was typically selected based on certain criteria after estimating the temperature-mortality relationship. However, the change point detected using this method is influenced by the smoothing of the function.
4. The permutation-based change-points detection procedure is introduced to test the significance of the detected change points under the CP-SISM. The previous work smoothed the nonparametric function and selected the change point that has minimum mortality or is based on some criteria, such as AIC or BIC. The permutation test is more powerful and robust compared to other tests/criteria-based analyses.

The remainder of this paper is organized as follows. In Section, the motivating data of this study is introduced. In Section, the proposed model is presented. A simultaneous procedure for estimating the proposed model while detecting and testing the significance of spatially dependent change points based on a permutation test is introduced in Section. In Section, several simulation studies are conducted. Section considers applying our unified method to South Korea's real data. Section includes discussion and conclusion.

Data and motivation

In our motivating data, non-accident mortality and weather variables, such as mean pressure, mean temperature, mean humidity, and time, were recorded daily from January 2000 to December 2007 for six major cities in South Korea (Seoul, Busan, Daegu, Incheon, Gwangju, and Daejeon). In total, the data comprise 2922 observations for each city. In addition, weekly data are obtained where daily weather variables were averaged such that each city has 417 observations. Because those cities are different in population size, the weekly non-accident mortality of each city is divided by the population size and multiplied by 1 million to obtain weekly nonaccident mortality per 1 million persons for each city. The numerical summary

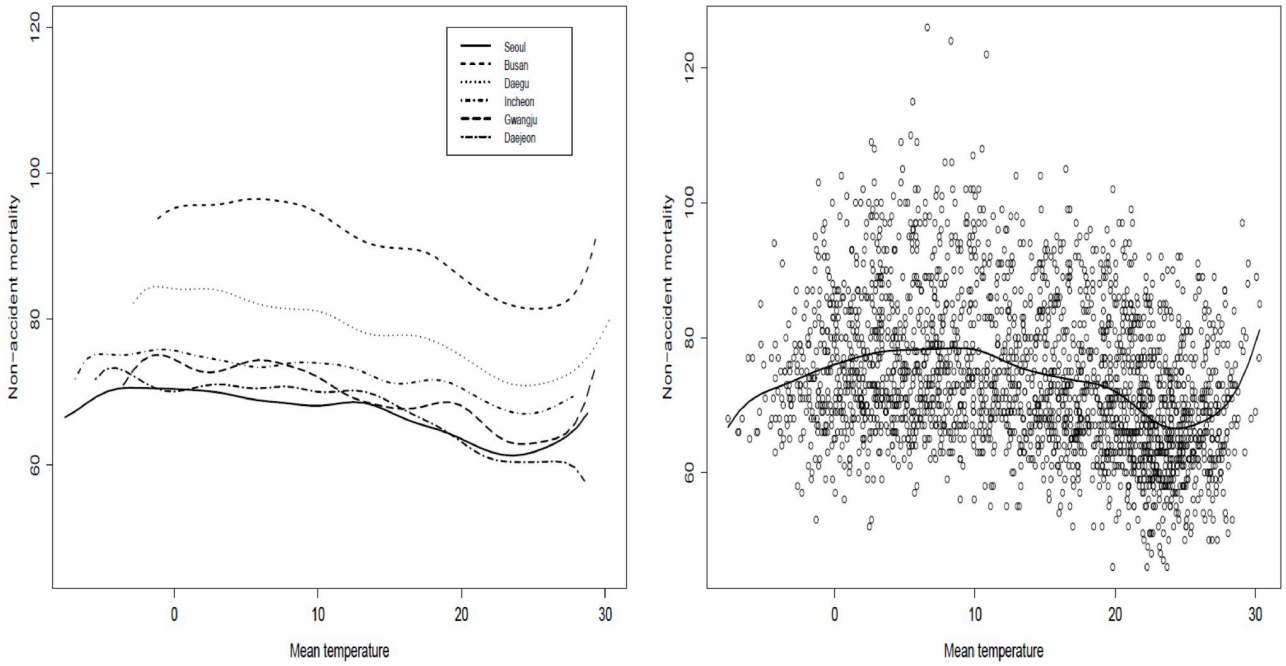

Fig 1. The aggregated smoothed mortality-temperature function of all cities along with scatter plot (a), and smoothed mortality-temperature functions of the six major areas in Korea (b).

https://doi.org/10.1371/journal.pone.0315413.g001

statistics of the weather variables of each city are presented in S1 Table of the supporting information.

Fig 1(a) shows a common change point of aggregated data of the six cities. Fig 1(b) reveals that the smoothed functions of non-accident mortality and temperature are similar in shape and show change points of all cities compared to each other. It shows there are possible change points in four cities (Seoul, Busan, Daegu, and Gwangju) and for the other two cities, it is not clear. That is because detecting a change point is affected by smoothing the unknown function, and change points commonly close to the boundaries where the smoothed function is located are not accurate. By focusing only on the interval that has a possible change point, Fig 2 shows that each city has a possible change point at some degree of temperature. These change points need to be studied simultaneously to see whether they are spatially dependent after adjusting the relationship by the weather variables. One important question here is whether these change points are statistically significant and/or significantly different from each other (i.e., spatially dependent). Hence, we study two cases by introducing a semiparametric model: a common change point (change points are not spatially dependent) and different change points over locations (spatially dependent).

Semiparametric change points single index spatial random effects model

Let Y_{is} be the ith observation at the sth city (location/region), and let x_{jis} be the ith value of the jth explanatory variable at city s, where $i = 1, \ldots, n$, $s = 1, \ldots, r$, and $j = 1, \ldots, p$. Here, n, r, and p denote the total number of observations, the number of locations, and the number of explanatory variables, respectively. Let $(\theta_s^1, \ldots, \theta_s^L)$ denote the possible multiple L change points at city s and $[x_{1is} - \theta_s^l]_+ = \max(0, x_{1is} - \theta_s^l)$ and $l = 1, \ldots, L$. We denote $f(\cdot)$ as the unknown mean

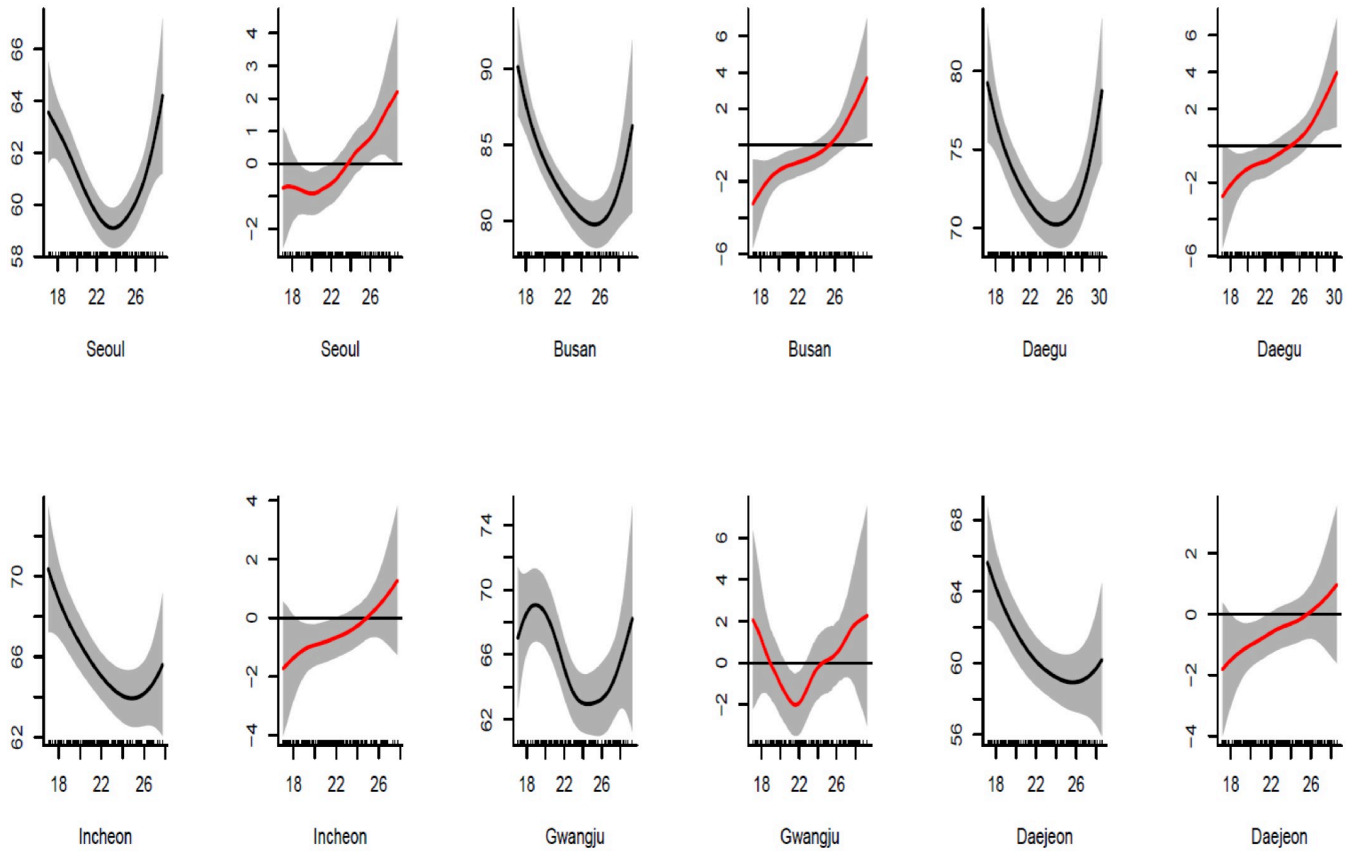

Fig 2. Spline smoothed mortality-temperature function of each city along with the smoothed derivative function. The black line represents the smoothed temperature-mortality function (the x-axis is the mean temperature and the y-axis is the smoothed mean temperature) and the red line is the derivative function of the temperature-mortality function (the x-axis is the mean temperature and the y-axis is the derivative of the mean temperature-mortality).

https://doi.org/10.1371/journal.pone.0315413.g002

function of the response variable. Let u_s be the spatial random effect associated with the sth city that follows a Gaussian process (GP) with covariance matrix $\boldsymbol{\Omega}$, and let ϵ_{is} be the random error associated with the ith observation at city s. We further denote a probability density/mass function of y_s as $p_d(y_s|\mu_s, u_s)$. The CP-SISM can be written as

$$\begin{aligned} y_{is}|\mu_s, u_s &\sim p_d(y_s|\mu_s, u_s), \\ \mu_s|u_s &= f(\beta_1 x_{1is} + \beta_{11}[x_{1is} - \theta_s^1]_+ + \ldots + \beta_{1L}[x_{1is} - \theta_s^L]_+ \\ &\quad + \beta_2 x_{2is} + \beta_3 x_{3is} + \ldots + \beta_p x_{pis}) + u_s, \\ u_s &\sim GP(0, \sigma_u^2 \boldsymbol{\Omega}), \end{aligned} \tag{1}$$

where

- the spatial effect, u_s ($s \in R^2$), follows a Gaussian stationary process with mean $\mathbf{0}$ for all s and a variance-covariance matrix depends only on the distance between any two locations s and $s + a$; $\text{cov}(u_s, u_{s+a}) = C(a)$ for all s, $a \in R^2$, where $C(\cdot)$ is a parametric covariance function, and a is the distance between two cities;
- $\boldsymbol{\beta} = (\beta_1, \beta_{11}, \ldots, \beta_p)$ represents the vector of the single index coefficient parameter and $\boldsymbol{\theta}_s = (\theta_s^1, \ldots, \theta_s^L)$ denotes the unknown parameters for multiple change points at city s;

- Given u_s and $\boldsymbol{\theta}_s$, $\mathbf{y}_s$ follows the Poisson distribution (Pois) with mean $E(\mathbf{y}_s|u_s, \boldsymbol{\theta}_s)$.

In matrix form, this model (1) can be written as

$$\begin{aligned} \mathbf{y}|\boldsymbol{\mu}, \mathbf{u} &\sim \text{Pois}(\boldsymbol{\mu}|\mathbf{u}, \boldsymbol{\alpha}), \\ \boldsymbol{\mu}|\mathbf{u}, \boldsymbol{\alpha} &= f(X(\boldsymbol{\theta})\boldsymbol{\beta}) + Z\mathbf{u}, \end{aligned} \tag{2}$$

where $X = [\mathbf{x}_1, (\mathbf{x}_1 - Z\boldsymbol{\theta})_+, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_p] \equiv X(\boldsymbol{\theta})$ is a $rn \times (p + L)$ matrix of regressors' values, $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\beta})^T$ is a $(p + L) \times 1$ vector of parameters, Z is a $nr \times r$ matrix of 1s, $\mathbf{u}$ is a vector of unobservable spatial correlated random effects, $\mathbf{u} \sim \text{MN}(\mathbf{0}, \sigma_u^2\boldsymbol{\Omega})$, where σ_u^2 is the variance of spatial effects and $\boldsymbol{\Omega}$ is a known parametric covariance function that depends on the distance between two cities. The random process is assumed to be stationary and isotropic, and the covariance between two cities depends on the distance between them.

Spatial Gaussian Processes provide a robust, flexible, and interpretable approach for spatial modeling, especially when dealing with continuous spatial variation and complex dependencies. Their adaptability, particularly in terms of covariance functions and Bayesian compatibility, makes them a superior choice in many contexts compared to SAR, CAR, or traditional kriging methods, which may impose more restrictive assumptions on the spatial data, [21–24].

More specifically, for our motivating real data, which has non-accident mortality as the response variable, we can write CP-SISM as the following:

$$\begin{aligned} \mathbf{y}_s|\boldsymbol{\mu}_s &\sim \text{Pois}(\boldsymbol{\mu}_s|u_s, \boldsymbol{\beta}, \boldsymbol{\theta}_s); \\ \boldsymbol{\mu}_s|u_s, \boldsymbol{\beta}, \boldsymbol{\theta}_s &= f(\beta_1\mathbf{x}_{1s} + \beta_{11}[\mathbf{x}_{1s} - \theta_s^1]_+ + \ldots + \beta_{1L}[x_{1is} - \theta_s^L]_+ \\ &\quad + \beta_2\mathbf{x}_{2s} + \beta_3\mathbf{x}_{3s} + \ldots + \beta_p\mathbf{x}_{ps}) + u_s. \end{aligned} \tag{3}$$

The unknown function, $f(\cdot)$, spatial effect, u_s, single index coefficients parameters, $\boldsymbol{\beta}$, and the vector of change points, $\boldsymbol{\theta}$, need to be estimated simultaneously and to be tested as to whether the change points are significant. The model parameters estimation needs a restriction on the single index coefficient parameters to fix the identifiability problem. A possible restriction is to set one of the parameters of $\boldsymbol{\beta}$ to be equal to 1 [25, 26] or to use $\|\boldsymbol{\beta}\| = 1$ [27–29].

This restriction prevents parameters from taking values that lead to indistinguishable outcomes, enabling the model to have a unique solution and thus be identifiable. Additionally, it reduces the model's complexity, preventing it from overfitting to noise in the data. This is particularly important in high-dimensional settings, as it stabilizes the estimation process by narrowing the parameter space. It also produces a simpler model, making it easier to interpret the impact of each coefficient, and helps the optimization algorithm converge more quickly and accurately, avoiding issues like local minima or divergence during estimation.

This model has several advantages: (1) It enables us to incorporate spatial effects into the model, (2) it enables us to detect multiple change points for each city, (3) it avoids the curse of the dimensionality problem by using the single index function, and (4) it is more flexible compared to the parametric models.

Change-point detection and testing

In this section, we propose a testing procedure to identify the significant spatially dependent change points. This procedure consists of an estimation step and a test step. These two steps are iterated until significant change points are detected if they exist. The estimation step for CP-SISM is based on an adjusted Monte Carlo Expectation Maximization (MCEM) algorithm. The EM algorithm consists of two steps: expectation (E-step) in which the spatial effects are estimated ($\mathbf{u} = u_1, u_2, \ldots, u_r$) and maximization (M-step) in which the variance of the spatial

effects (σ_u^2) is estimated. The vector of the coefficient parameters ($\boldsymbol{\beta} = \beta_1, \ldots, \beta_p$) is estimated using the Ichimora method, and the $f(index)$ function is estimated using a smoothing method, such as the kernel method.

The estimation of $f(\cdot)$ using the Ichimura method is performed as follows:

Step 0: For a given estimate of the index coefficient vector $\boldsymbol{\beta}$, we compute the single-index values $Z_i = X_i\boldsymbol{\beta}$, where $i = 1, \ldots, n$.

Step 1: The unknown function $f(\cdot)$ is estimated using *kernel smoothing*. Specifically, for any value z, $f(z)$ is estimated as:

$$\hat{f}(z) = \frac{\sum_{i=1}^{n} K_h(z - Z_i)y_i}{\sum_{i=1}^{n} K_h(z - Z_i)},$$

where:$K_h(\cdot)$ is a kernel function with a bandwidth h, y_i are the observed responses, and $Z_i = \boldsymbol{\beta}^\top X_i$ are the single-index values. This approach smooths the observed y values as a function of the single-index Z, providing an estimate of $f(\cdot)$.

Step 2: The estimation of $f(\cdot)$ and $\boldsymbol{\beta}$ is performed iteratively. After updating $\boldsymbol{\beta}$ using optimization techniques, $f(\cdot)$ is re-estimated based on the updated single-index values until convergence is achieved.

To estimate a change point, a grid search is used. At each possible change point, the EM algorithm is run to estimate the spatial effects and model parameters, and the test procedure is used to see whether it is a significant change point. So the order of the estimation is as follows: at each possible change point, the EM algorithm is used to estimate the spatial effects and variance of spatial effects, and then model parameters and the unknown function are estimated using the Ichimura method. For each possible change point, the sum of the squared residuals is calculated and the change point associated with the minimum sum of squared residuals is selected and is then tested to determine whether it is significant based on the calculated p-value of the permutation test.

Estimation step

The estimation step of CP-SISM is based on an adjusted MCEM algorithm. The EM algorithm consists of an expectation (E-step) and a maximization (M-step). Incorporating the Monte Carlo step into the EM algorithm gives the MCEM algorithm, which is commonly used in the generalized linear mixed models estimation [30–35].

Our proposed model, CP-SISM, has the following complete-data log-likelihood form:

$$\log f_{\mathbf{y}_s, u_s}(\mathbf{y}_s, u_s|\boldsymbol{\mu}_s, \sigma_u^2, \Omega) = \log f_{\mathbf{y}_s}(\mathbf{y}_s|\boldsymbol{\mu}_s, u_s) + \log f_{u_s}(u_s|\sigma_u^2\Omega), \tag{4}$$

where $\mathbf{y}_s \sim \text{Pois}(\boldsymbol{\mu}_s|\mathrm{u}_s)$, $u_s \sim GP(0, \sigma_u^2\Omega)$, $\boldsymbol{\mu}_s|u_s = f(X_s\boldsymbol{\beta}) + u_s$, $\Omega = \text{Cov}(\mathrm{u}_s, \mathrm{u}_{s+a}) = \exp(||\mathrm{a}||^2/\rho_\mathrm{u})$ for all s, $a \in R^2$, a is the distance between two cities s and $s + a$, and ρ_u is the dependence range.

In the E-step of the MCEM algorithm for our model estimation, there is no closed form available. Hence, random samples are generated from the full conditional distribution of $\mathbf{u}$ using Bayesian MCMC. The single-component Metropolis-Hastings (M-H) algorithm is used, i.e., a single component is updated at each iteration, say the sth component, u_s. Selecting a proposal function is essential in the M-H algorithm. Because the spatial random effects are correlated, we propose generating candidate values from the conditional normal distribution $N(\bar{\gamma}, \sigma_0^2\bar{\Omega})$. The following illustrates the derivation of the conditional normal distribution. Let

$\mathbf{v} = (v_1, v_2, \ldots, v_n) = (v_1, \mathbf{v_2})^T$ have multivariate normal distribution with mean $\boldsymbol{\gamma} = (\gamma_1\ \gamma_2)^{\mathrm{T}}$ and variance-covariance matrix $\sigma_0^2\Omega$, where

$$\Omega = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

The distribution of v_1, given that $\mathbf{v_2} = \mathbf{a}$, is a multivariate normal distribution $(v_1|\mathbf{v_2} = \mathbf{a}) \sim N(\bar{\gamma}, \sigma_0^2\bar{\Sigma})$, where $\bar{\gamma} = \gamma_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{a} - \gamma_2)$ and variance-covariance matrix $\bar{\Sigma} = \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. So the conditional normal distribution, $N(\bar{\gamma}, \sigma_0^2\bar{\Sigma})$, is our proposal distribution, where σ_0^2 is the proposal variance of the correlated spatial random effects.

Similarly, given the other spatial random effects, we obtain the conditional normal distribution of u_s as $N(\bar{\gamma}, \sigma_u^2\bar{\Omega})$, where $\bar{\gamma} = \Omega_{12}\Omega_{22}^{-1}(\mathbf{a})$ and $\bar{\Omega} = \sigma_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$. As a result, the acceptance probability, in E-step, can be written as

$$\min\left[\frac{f(\mathbf{y}_s|u_s^*, \boldsymbol{\mu}_s)f_u(u_s^*|\bar{\gamma}, \sigma_u^2\bar{\Omega})}{f(\mathbf{y}_s|u_s, \boldsymbol{\mu}_s)f_u(u_s|\bar{\gamma}, \sigma_u^2\bar{\Omega})}, 1\right], \tag{5}$$

where $f_{u_s}(u_s|\bar{\gamma}, \sigma_u^2\bar{\Omega})$ is the conditional distribution of u_s that is given all the other spatial random effects.

In M-step, given spatial effects and candidate spatial change points, $\sum \log f(\mathbf{u}|\sigma_u^2\Omega)$ is maximized to obtain $\hat{\sigma}_u^2$, estimate $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, and smooth the function $f(\cdot)$ to obtain $\hat{f}(\cdot)$. Then E-step and M-step are iterated until the convergence is achieved.

Testing step

In this section, we explain how to conduct the testing procedure by connecting a nonparametric Poisson regression with a single index nonparametric function $f(\cdot)$ that can estimate the link function as well. In Poisson regression with an unknown function $m(\cdot)$ and a link function $g(\cdot)$, we can express the model as

$$\begin{aligned} g\{E(y|\mu)\} &= m\{X(\boldsymbol{\theta})\boldsymbol{\beta}\} + Zu; \\ E(y|\mu) &= g^{-1}[m\{X(\boldsymbol{\theta})\beta\} + Zu]; \\ &= g^{-1}[m\{X(\boldsymbol{\theta})\beta\}] \times g^{-1}(Zu); \\ &= f\{X(\boldsymbol{\theta})\beta\} \times g^{-1}(Zu); \\ &\approx [f\{X(\boldsymbol{\theta})\beta\} + c] \times [g^{-1}(0) + \{g^{-1}(0)\}'Zu + O(\|Zu\|)]; \\ &= f\{X(\boldsymbol{\theta})\boldsymbol{\beta}\}g^{-1}(0) + c\{g^{-1}(0)\}'Zu + O(\|Zu\|). \end{aligned} \tag{6}$$

Because $g^{-1}(0)$ and $c\{g^{-1}(0)\}'$ are both constants, they can be merged to the unknown function $f(\cdot)$ and random variable u. Hence, we can develop the testing procedure under the following approximated model,

$$\mathbf{y} \approx f(X(\boldsymbol{\theta})\boldsymbol{\beta}) + Zu + \epsilon \tag{7}$$

where $\epsilon = \mathbf{y} - \boldsymbol{\mu}$. Hence, our permutation testing procedure is developed under this approximation.

The multiple spatially dependent change-point candidates in cities, $\boldsymbol{\theta} = (\theta_s^1, \theta_s^2, \ldots, \theta_s^L)$, $s = 1, \ldots, r$, are tested to determine whether they are significant based on our permutation-based testing approach described as follows.

Under the null hypothesis of no change points, CP-SISM can be written as

$$\begin{aligned} \mathbf{y}|\boldsymbol{\mu}^{(\mathbf{0})} &\sim \text{Pois}(\boldsymbol{\mu}^{(0)}|\boldsymbol{\beta}, \mathbf{u}), \\ \boldsymbol{\mu}^{(0)} &= f(\beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \ldots + \beta_p\mathbf{x}_p) + \mathbf{u}, \\ \epsilon^{(0)} &= \mathbf{y} - \boldsymbol{\mu}^{(0)}. \end{aligned} \tag{8}$$

Under the alternative hypothesis with $\boldsymbol{\theta}$ vector of change points, SCP-SIM takes the following form:

$$\begin{aligned} \mathbf{y}|\boldsymbol{\mu}^{(\mathbf{1})} &\sim \text{Pois}(\boldsymbol{\mu}^{(1)}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{u}), \\ \boldsymbol{\mu}^{(1)} &= f(\beta_1\mathbf{x}_1 + \beta_{11}[\mathbf{x}_1 - Z\boldsymbol{\theta}^1]_+ + \ldots + \beta_{1L}[\mathbf{x}_1 - Z\boldsymbol{\theta}^L]_+ + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \ldots \\ &\quad +\beta_p\mathbf{x}_p) + \mathbf{u}, \\ \epsilon^{(1)} &= \mathbf{y} - \boldsymbol{\mu}^{(1)}. \end{aligned} \tag{9}$$

The test statistic is based on the ratio of the residuals of the original data,

$$T_{\mathbf{y}_{(0)}} = \frac{[\hat{\epsilon}^{(0)}_{\mathbf{y}_{(0)}}]'[\hat{\epsilon}^{(0)}_{\mathbf{y}_{(0)}}]}{[\hat{\epsilon}^{(1)}_{\mathbf{y}_{(0)}}]'[\hat{\epsilon}^{(1)}_{\mathbf{y}_{(0)}}]}, \tag{10}$$

where $\hat{\epsilon}^{(0)}_{\mathbf{y}_{(0)}}$ and $\hat{\epsilon}^{(1)}_{\mathbf{y}_{(0)}}$ denote the residuals under the null and alternative hypotheses of the actual data $\mathbf{y}_0$. Permutation-based p-value can be calculated and the candidate spatially dependent change points are declared significant if p-value $<\alpha$, where α is the significant level.

When multiple change points are considered, $L > 1$, H_0: $L_0 = 0$ versus H_1: $L_1 = L$, where L is the possible number of change points, is tested. If H_0 is rejected, we then test H_0: $L_0 = 1$ versus H_1: $L_1 = L$; otherwise, we test H_0: $L_0 = 0$ versus H_1: $L_1 = L - 1$ until we reach testing H_0: $L_0 = l$ versus H_1: $L_1 = l + 1$. For the last two hypotheses, if H_0, is rejected, then the number of significant change points declared is $l + 1$, otherwise, it is l.

Simulation studies

Three simulation studies are conducted to evaluate the performance of our approach in detecting and testing change points. We assume that the number of change points is unknown and fixed. We first determine the potential maximum number of change points, K_{max}, and then conduct the permutation test to identify the number of significant change points. We consider the following three cases: (1) when there are no change points, (2) when there is only one change point, and (3) when there are two change points. Using the likelihood ratio test proposed by [11], we determined K_{max}. We can treat the first case with zero significant changes out of K_{max} as a type I estimated error. The second case is considered to have occurred when one change point is significant out of K_{max} and the other change points are not significant. The third case is considered to have occurred when the two change points are significant and the other change points are not significant.

Simulation Study 1: No change points—Type I error

Type I error is studied by simulating 100 data sets from the CP-SISM with no change points (the model under the null hypothesis) that takes the form:

$$\begin{aligned} y_{is} &\sim \text{Pois}(\mu_{is}|\boldsymbol{\beta}, u_s), \\ \mu_{is} &= f(\beta_1 x_{1is} + \beta_2 x_{2is} + \beta_3 x_{3is}) + u_s, \\ & i = 1, 2, \ldots, n \text{ and } s = 1, 2, \ldots, r, \end{aligned} \tag{11}$$

with six locations ($r = 6$) and 100 observations at each location ($n = 100$). We set the true parameters as $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3) = (-0.5, 1, 1)$, $\sigma_u = 1$, so the mean function is equal to $\boldsymbol{\mu}_s = f(X\boldsymbol{\beta}) + u_s = (-0.5x_1 + x_2 + x_3)^2 + u_s$. In this setting, there is no change point. The permutation test is used to detect any significant change point and it is found that the null hypothesis is rejected 6 times out of the 100. This means that the Type I error of the test is maintained approximately at 5%.

Simulation Study 2: A single change point at each city

In this section, two cases are considered: (1) there is one common change point for all cities, and (2) there are different change points for cities.

One common change point for all cities. One hundred data sets are simulated from the proposed model (CP-SISM), in which

$$\begin{aligned} y_{is} &\sim \text{Pois}(\mu_{is}|\boldsymbol{\beta}, \theta, u_s), \\ \mu_{is} &= f(\beta_1 x_{1is} + \beta_{11}[x_{1is} - \theta]_+ + \beta_2 x_{2is} + \beta_3 x_{3is}) + u_s, \\ & i = 1, 2, \ldots, n \text{ and } s = 1, 2, \ldots, r, \end{aligned} \tag{12}$$

with six locations ($r = 6$) and 100 observations at each location ($n = 100$). Three explanatory variables (x_1, x_2, and x_3) are generated from Uniform(π, 2π). We set true parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_{11}) = (-0.5, 1, 1, 1)$, $(\theta, \sigma_u) = (4.7, 1)$, and the mean function $\boldsymbol{\mu}_s = f(X(\boldsymbol{\theta})\boldsymbol{\beta}) + u_s = (-0.5x_1 + [x_1 - 4.7]_+ + x_2 + x_3)^2 + u_s$. In this setting, there is a common change point at $\theta_s = \theta = 4.7$ for all s ($s = 1, 2, \ldots, r$). Here, β_2 is set to 1 to fix the identifiability problem. Based on the mean squared error (MSE), and mean, median and inter-quartile range (IQR) of the estimates, the estimation would be evaluated. In addition, the proportion of the significant detected change points is calculated.

We set the dependence range $\rho_u = 2$, and the variance of the spatial effects $\sigma_u = 1$. The domain of $[0, 3] \times [0, 3]$ is used in this simulation study because the range of the distance between spatial locations of latitude and longitude in the motivating data set is found to be about 2. $\mathbf{y}_s|\boldsymbol{\mu}_s$ is generated from Poisson distribution with mean $\boldsymbol{\mu}_s|u_s$. The reason for not using a large value of σ_u in the simulation is to ensure we do not obtain a negative mean value, where the response variable has the Poisson distribution.

Table 1 shows that the mean estimates of all the parameters are close to the true values for all the parameters and change points.

To obtain the empirical coverage probability for the model parameters and the change point, 500 data sets are simulated based on the setting that is described above, and the model parameters and the change point are estimated for each simulated data set. Then, 10,000 random samples of size 30, with replacement, are selected from each parameter estimate (the model parameters and change point) and a 95% confidence interval is calculated for each parameter, for each sample. The coverage probability of each parameter is estimated by

Table 1. Results for 500 simulated data sets from CP-SISM with one common change point; the mean, median, MSE, 95% confidence interval, and empirical coverage probability of the model parameters and change point.

	True	Mean	Median	MSE	95% CI	Coverage probability
β_1	-0.5	-0.5174	-0.4984	0.024	(-0.547, -0.488)	94.27%
β_3	1	1.0056	1.0026	0.003	(0.9949, 1.0103)	94.07%
β_{11}	1	1.0300	1.0172	0.029	(0.9925, 1.0419)	91.02%
θ	4.7	4.693	4.7000	0.054	(4.6731, 4.7137)	95.51%
σ_u	1	0.9964	0.9620	0.159	(0.9614, 1.0314)	94.53%

https://doi.org/10.1371/journal.pone.0315413.t001

calculating the proportion of the confidence intervals that contain the true parameters. The results are reported in Table 1, which shows the confidence intervals achieve the near nominal coverage probability.

Different change points for cities. Similar to the simulation in Study 1, 100 data sets are simulated from the following model:

$$\begin{aligned} y_{is} &\sim \text{Pois}(\mu_{is}|\boldsymbol{\beta}, \theta_s, u_s), \\ \mu_{is} &= f(\beta_1 x_{1is} + \beta_{11}[x_{1is} - \theta_s]_+ + \beta_2 x_{2is} + \beta_3 x_{3is}) + u_s, \end{aligned} \tag{13}$$

$$i = 1, 2, \ldots, n \text{ and } s = 1, 2, \ldots, r,$$

with six locations ($r = 6$), 100 observations at each location ($n = 100$). Three explanatory variables (x_1, x_2, x_3) are generated from Uniform(3, 4). We set true parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_{11}) = (1, 0.3, 0.3, 3)$, and the mean function $\mu_s = f(X(\boldsymbol{\theta})\boldsymbol{\beta}) + u_s = (x_1 + 3[x_1 - \theta_s]_+ + 0.3x_2 + 0.3x_3)^2 + u_s$. The data are generated such that every two locations share the same change point, three different spatial change points in total. The first two locations have a change point at 3.2 ($\theta_1 = \theta_2 = 3.2$), the second two locations have change points at 3.5 ($\theta_3 = \theta_4 = 3.5$), and the last two locations each have a change point at 3.8 ($\theta_5 = \theta_6 = 3.8$). The variance of the spatial effects is 1, $\sigma_u = 1$. The [0, 3] × [0, 3] domain is used in this simulation study. Three cases are considered for the dependence range ($\rho_u = 0.5$, 1, and 2). Here, $\rho_u = 0.5$ means there is not much dependence, and $\rho_u = 2$ means a high dependence range. $\mathbf{y}_s|\boldsymbol{\mu}_s$ is generated from the Poisson distribution with mean $\boldsymbol{\mu}_s|u_s$. Fig 3(a) shows a random simulated data set based on this setting. Under this setting, it is found that the spatial variance estimate is over-estimated, so a penalty value is used, λ, in the M-step of the proposed estimation algorithm.

Fig 3(b) shows the average of AIC at each value of λ. It reveals that the optimal value is about 1.9. One hundred data sets are generated from this setting, and using the optimal value of λ, the MSE, mean, median, and IQR of the estimates are calculated to evaluate the estimating approach. The permutation-based test is used to test the significance of the detected change points and the proportion of the significant detected change points is calculated as well. Table 2 shows the results of the 100 simulated data sets. It shows that the performance of the proposed model in detecting change points works well. The model parameter estimates are close to the true values and have quite small standard error and MSE. The model parameter estimates and detection of change points under different values of dependence range, ρ_u, are comparable. The proportions of significant detected change points for the different values of the dependent range are 98% for $\rho_u = 0.5$, 97% for $\rho_u = 1$, and 97% for $\rho_u = 2$.

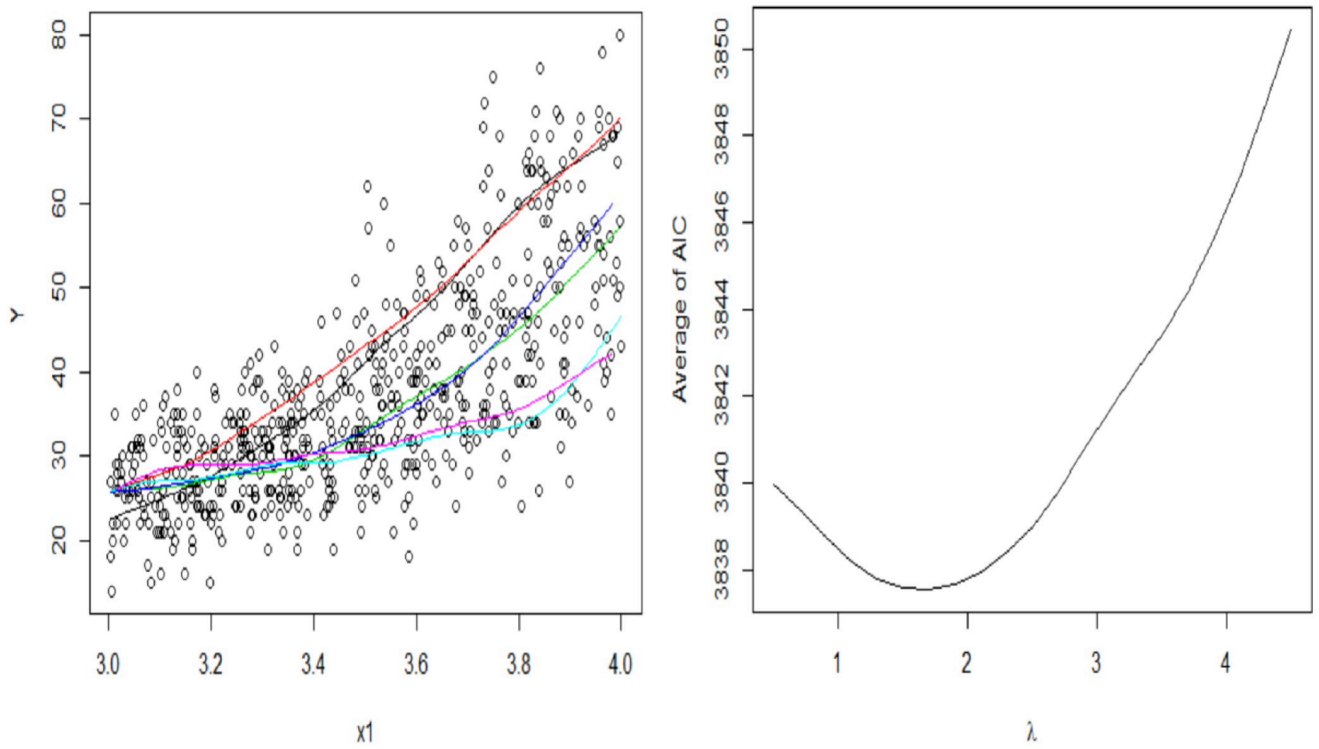

Fig 3. Simulated data set of six locations, every two locations has equal change points (a), and average AIC versus the penalty value, λ (b).

https://doi.org/10.1371/journal.pone.0315413.g003

Table 2. Results of 100 simulated data sets: MSE, mean, median, and standard error of the model parameters and change points estimates for different values at different dependence range, $\rho_u = 0.5, 1, 2$.

		True	Mean ± SE	MSE	Median	IQR
$\rho = 0.5$	β_2	0.3	0.31 ± 0.021	0.033	0.27	0.08
	β_3	0.3	0.30 ± 0.043	0.042	0.30	0.10
	β_{11}	3	3.09 ± 0.067	0.130	2.75	0.15
	θ_1	3.2	3.21 ± 0.001	0.002	3.20	0.05
	θ_2	3.5	3.51 ± 0.007	0.004	3.55	0.10
	θ_3	3.8	3.78 ± 0.005	0.007	3.80	0.05
	σ_u	1	1.03 ± 0.051	0.122	0.93	0.15
$\rho = 1$	β_2	0.3	0.29 ± 0.020	0.032	0.26	0.09
	β_3	0.3	0.30 ± 0.046	0.042	0.27	0.08
	β_{11}	3	2.89 ± 0.065	0.130	2.67	0.13
	θ_1	3.2	3.23 ± 0.001	0.002	3.20	0.05
	θ_2	3.5	3.49 ± 0.002	0.004	3.45	0.10
	θ_3	3.8	3.73 ± 0.002	0.007	3.75	0.05
	σ_u	1	0.97 ± 0.041	0.124	0.92	0.15
$\rho = 2$	β_2	0.3	0.26 ± 0.027	0.033	0.24	0.10
	β_3	0.3	0.25 ± 0.095	0.046	0.24	0.11
	β_{11}	3	2.66 ± 0.105	0.134	2.33	0.14
	θ_1	3.2	3.16 ± 0.015	0.005	3.15	0.10
	θ_2	3.5	3.45 ± 0.012	0.006	3.40	0.10
	θ_3	3.8	3.74 ± 0.008	0.010	3.70	0.05
	σ_u	1	0.98 ± 0.128	0.138	0.94	0.19

https://doi.org/10.1371/journal.pone.0315413.t002

Simulation Study 3: Two change points

One hundred and fifty data sets are simulated from the proposed model (CP-SISM),

$$\begin{aligned} y_{is} &\sim \text{Pois}(\mu_{is}|\boldsymbol{\beta}, \theta_s, u_s), \\ \mu_{is} &= f(\beta_1 x_{1is} + \beta_{11}[x_{1is} - \theta_1]_+ + \beta_{12}[x_{1is} - \theta_2]_+ + \beta_2 x_{2is}) + u_s, \end{aligned}$$

$$i = 1, 2, \ldots, n \text{ and } s = 1, 2, \ldots, r,$$

with six locations ($r = 6$) and 100 observations at each location ($n = 100$). Two explanatory variables (x_1 and x_2) are generated from Uniform(π, 3π). We set true parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_{11}, \beta_{12}) = (1, 1, -2, 1.5)$, $(\theta_1, \theta_2, \sigma_u) = (4.5, 7.5, 1)$, and the mean function $\boldsymbol{\mu}_s = f(X(\boldsymbol{\theta})\boldsymbol{\beta}) + u_s = (x_1 + [x_1 - 4.5]_+ + [x_1 - 7.5]_+ + x_2)^2 + u_s$. In this setting, there are two change points at $\theta_1 = 4.5$ and $\theta_2 = 7.5$ for all locations. Here, β_2 is set to 1 to fix the identifiability problem. Based on the mean, median, and MSE of the estimates and 95% confidence interval of the model parameters, the estimation would be evaluated. Table 3 shows that in the estimation based on the criteria that are used, the model parameters are well estimated.

Real data application

In this section, our approach is applied to our motivating data. Non-accidental mortality (ICD-10 codes A00-R99) data are obtained from Statistics Korea and historical weather data, such as daily average temperature, pressure, and humidity, are obtained from the Korea Meteorological Administration. Non-accidental mortality (excluding deaths related to accidents) is chosen because it has been widely used in previous studies. The non-accident mortality and weather variables were recorded daily from January 2000 to December 2007 for six major cities in South Korea: Busan, Daegu, Daejeon, Gwangju, Incheon, and Seoul. The total number of records is 2,922 days with 668,583 deaths. The weekly data are also created from this daily data, which resulted in 417 observations at each city. The latitude and longitude in this motivating data set are further explained in S1 Table of the supporting information.

In previous studies, a change point was estimated for each city separately and a common change point was considered [16, 18, 20, 36]. In addition, testing the change point was not conducted.

Our goals are simultaneously (1) estimate the relationship between the weekly non-accident mortality ($\mathbf{y}$) and mean temperature ($\mathbf{x}_1$), adjusting for other covariates such as mean humidity ($\mathbf{x}_2$), mean pressure ($\mathbf{x}_3$), and month as a factor ($\mathbf{x}_4$); (2) to detect the possible spatially dependent change points in temperature of each city; and (3) to test whether the detected spatial change points are significant by using the proposed permutation-based test. In our weekly motivating data, we have four explanatory variables ($p = 4$) and 417 observations for each city ($n = 417$).

Table 3. Results for 150 simulated data sets from CP-SISM with two change points; the mean, median, MSE, and 95% confidence interval of the model parameters and change point.

	True	Mean	Median	MSE	95% CI
β_1	1	0.991	0.990	0.092	(0.943, 1.038)
β_{11}	-2	-2.002	-1.986	0.077	(-2.047, -1.959)
β_{12}	1.5	1.507	1.497	0.077	(1.480, 1.536)
θ_1	4.5	4.527	4.500	0.033	(4.498, 4.557)
θ_2	7.5	7.471	7.500	0.034	(7.439, 7.504)
σ_u	1	1.112	1.095	0.220	(1.045, 1.178)

https://doi.org/10.1371/journal.pone.0315413.t003

Table 4. Parameter estimates, standard errors, change points detected, p-values, and R^2 of CP-SISM and the aggregated data model, assuming there is one common change point.

	Aggregated Model	CP-SISM
$\hat{\beta}_1$	-1.769±0.035	-1.334±0.034
$\hat{\beta}_{11}$	2.439±0.036	1.782±0.033
$\hat{\beta}_2$	.0963±0.001	-0.145±0.0003
$\hat{\beta}_4$	-0.4852±0.018	0.319±0.0048
$\hat{\sigma}_u$	—	29.99
θ±$SE(\theta)$	22 ± 0.124	22 ± 0.111
p-value	0.019	0.009
R^2	0.22	0.69

https://doi.org/10.1371/journal.pone.0315413.t004

Detecting and testing a common change point

The proposed model, with a common change point of all cities $\theta_s = \theta$, has the form

$$\begin{aligned} \mathbf{y}_s &\sim \text{Pois}(\boldsymbol{\mu}_s|\boldsymbol{\beta}, \theta, u_s), \\ \boldsymbol{\mu}_s &= f(\beta_1 x_1 + \beta_{11}[x_1 - \theta]_+ + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4) + u_s, \end{aligned} \tag{14}$$

and with no common change point, it takes the form

$$\begin{aligned} \mathbf{y}_s &\sim \text{Pois}(\boldsymbol{\mu}_s|\boldsymbol{\beta}, \theta, u_s), \\ \boldsymbol{\mu}_s &= f(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4) + u_s, \qquad s = 1, 2, \ldots, r. \end{aligned} \tag{15}$$

One common change point is detected. We then test whether this detected change point is significant. The results are compared to the case of aggregating the data of all cities by ignoring spatial effects.

Table 4 shows that the proposed model, CP-SISM, fits the data better than the aggregated model does, in which the R^2 (=0.69) of CP-SISM is much higher than R^2 (=0.22) of the aggregated data model is. However, the same change point value ($\theta = 22^oC$) is detected and found to be significant, but the p-value of the proposed model is smaller. The standard error and confidence intervals of the parameters and change point are calculated using a permutation approach as follows:

Step 1. A sample of observations from each city (with replacement) is randomly selected.

Step 2. The model parameters and the change point are estimated.

1. Step 1—Step 2 are repeated 500 times and the standard error and confidence interval are calculated for each parameter and change point.

It is found that the change point estimate of the proposed model has a smaller standard error compared to that of the aggregated model. Fig 4 shows that detecting the change point by smoothing the unknown function does not give an accurate value. The change point by smoothing the unknown function is about 24^oC. However, based on the permutation test, the change point of the proposed model is about 22^oC. Fig 4(a) shows the smoothed function for aggregated data and Fig 4(b) shows the smoothed functions for the six cities from the proposed model. The smoothed function of the aggregated data is wigglier compared to the smoothed functions of the proposed model. Fig 5(b) shows that the highest mortality is for Busan and the lowest mortality function is for Gwangju and Seoul.

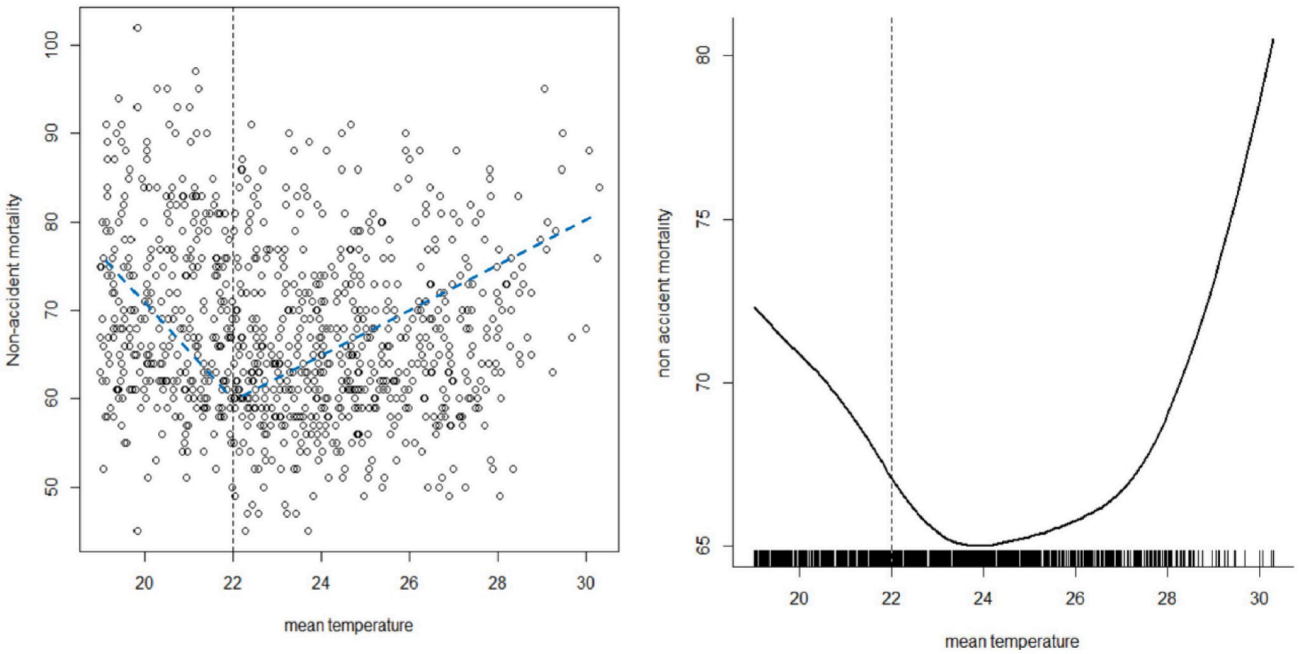

Fig 4. Scatter plot along with the detected common change point (a), Spline smoothed mortality-temperature function of all the six cities (b).

https://doi.org/10.1371/journal.pone.0315413.g004

Regarding the single index coefficient estimates and their standard errors, Table 4 shows that the standard errors estimated by the permutation method for the proposed model are smaller than those of the aggregated model. We also noticed that some of the coefficients are different in the sign. For both models, the coefficient of the mean pressure is set to 1 to fix the identifiability problem.

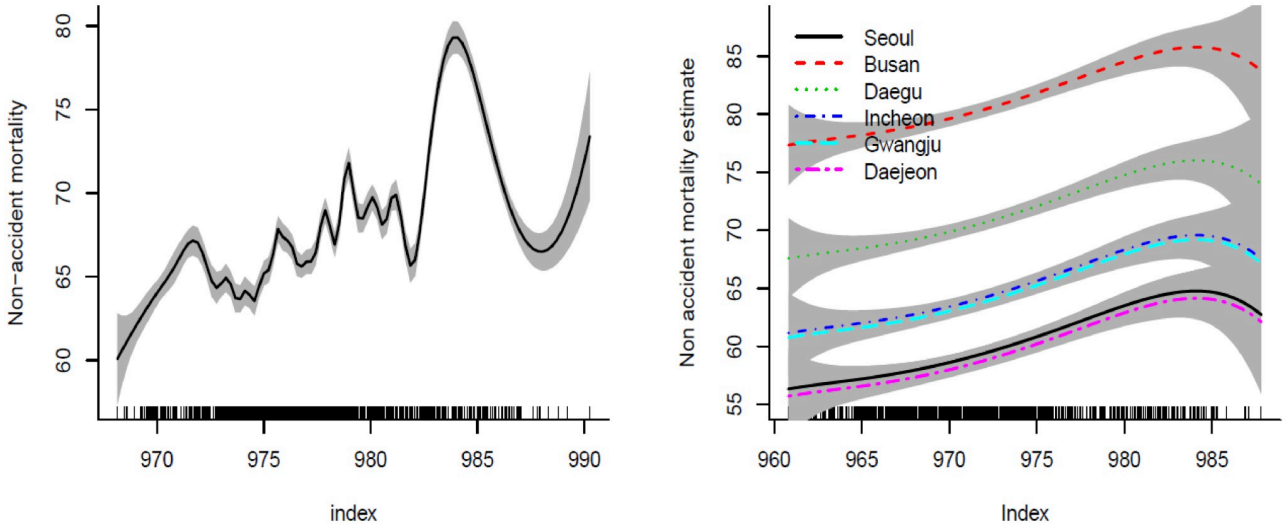

Fig 5. Smoothed mortality function of the aggregated data of the six cities (a) and the smoothed mortality functions from the proposed model (b).

https://doi.org/10.1371/journal.pone.0315413.g005

Table 5. Detected change points in the two cases: No spatial effects considered (change point is detected and tested for each city separately), and spatial effect considered in our proposed model, CP-SISM (change points are detected and tested simultaneously).

	No spatial		CP-SISM
	$\hat{\theta}_s \pm \boldsymbol{se}(\hat{\theta}_s)$	p-value	$\hat{\theta}_s \pm \boldsymbol{se}(\hat{\theta}_s)$
Busan	23.2±0.131	0.371	23.2±0.112
Incheon	22.8±0.132	0.148	22.8±0.122
Seoul	22.6±0.136	0.019	22.4±0.124
Daegu	22.6±0.133	0.059	22.6±0.121
Daejeon	22.6±0.142	0.029	22.6±0.109
Gwangju	22.4±0.131	0.049	22.4±0.113
			R^2 = 0.73 and p-value = 0.000

https://doi.org/10.1371/journal.pone.0315413.t005

Detecting and testing various change points

The proposed model has the following form:

$$\begin{aligned} \mathbf{y}_s &\sim \text{Pois}(\boldsymbol{\mu}_s|\boldsymbol{\beta}, \theta_s, u_s), \\ \boldsymbol{\mu}_s &= f(\beta_1 x_1 + \beta_{11}[x_1 - \theta_s]_+ + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4) + u_s, \qquad s = 1, 2, \ldots, r. \end{aligned} \tag{16}$$

Simultaneously, the model is estimated, the spatially dependent possible change points are detected, and the detected change points are tested to determine whether they are significant. The results are compared to the case of detecting a change point in each city separately and then tested to determine whether it is significant.

Table 5 shows the change points that are detected in case of no spatial effects (each city is analyzed separately) and the proposed model, CP-SISM, along with the standard errors and *p*-values. It is found that the change points that are detected and tested simultaneously are comparable to the no spatial effects case. However, the parameter estimates of the CP-SISM have smaller standard errors.

Under the CP-SISM, the smallest change point value is for Seoul and Gwangju (22.4), and the highest change point value is for Busan (23.2). In addition, for the no spatial effects case, one can see that three of the cities have insignificant change points (Incheon, Busan, and Daegu). For the CP-SISM, R^2 is improved (R^2 = 0.73) compared to the one common change point case (R^2 = 0.69). The improvement is not significant because the detected change points are close and close to the common change point value except Busan city change point which has a higher change point compared to the other cities. As a result, the difference between the two models' R^2 values is not big.

Fig 6(a) compares the change points detected in each city under the CP-SISM and under the case of detecting a change point in each city separately. To check whether these change points of the CP-SISM are different, the 95% confidence interval for each detected change point is calculated using the permuted standard error and shown in Fig 6(b). It reveals the confidence intervals of the detected change points overlap, except for Busan city. This explains why there is not much difference between the two cases: one common change point and the spatially dependent change points case. These two cases are also compared in terms of the model parameter estimates and the results are summarized in Table 6. It shows they have comparable parameter estimates and standard errors, as well as comparable estimate values of spatial effects. It shows that the smallest spatial random effects are of Seoul and Daejeon and the

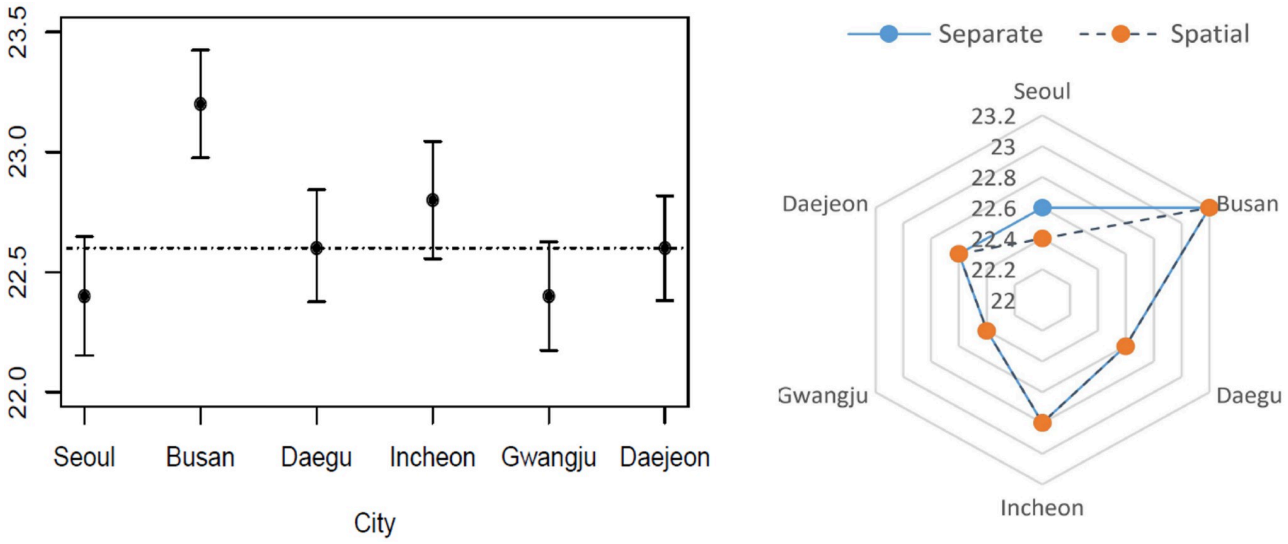

Fig 6. The 95% confidence interval for each change point (a) and a radar plot of the significant detected change points using the proposed model (CP-SISM) and no spatial random effects considered case (b).

https://doi.org/10.1371/journal.pone.0315413.g006

highest is of Busan, which is much higher compared to the other cities. Busan has the highest change point value and the highest mortality.

Discussion and conclusion

A semiparametric regression model (CP-SISM) is introduced to simultaneously estimate the nonlinear temperature-mortality relationship, detect spatially dependent change points, and test to determine whether they are significant based on a permutation-based test, and a unified method is proposed. Simulation studies are conducted for two cases: change points are spatially independent and change points are spatially dependent. Simulation studies showed that our approach works well in estimating, detecting, and testing spatial change points simultaneously.

The advantages of our proposed approach are demonstrated using epidemiology data on mortality and temperature, as well as other weather variables that were collected daily from six

Table 6. Parameter estimates and standard errors, spatial effect estimates of the proposed model (CP-SISM) in case one common spatial change point is assumed, θ, and in case different spatially-dependent change points are assumed, $\theta = (\theta_1, \ldots, \theta_6)$.

		CP-SISM	CP-SISM
		A common change point	Spatially-dependent change points
Parameter estimates	$\hat{\beta}_1$	-1.334± 0.035	-1.436± 0.032
	$\hat{\beta}_{11}$	1.782± 0.036	2.104± 0.031
	$\hat{\beta}_2$	-0.145±0.001	-0.145± 0.001
	$\hat{\beta}_4$	0.319± 0.018	0.2921±0.011
Spatial effects	Busan	13.28	13.99
	Incheon	-2.91	-2.18
	Seoul	-7.73	-7.11
	Daegu	3.53	4.30
	Daejeon	-8.35	-7.56
	Gwangju	-3.29	-2.73

https://doi.org/10.1371/journal.pone.0315413.t006

major cities in South Korea. It is found that cities have close change points, except Busan city, which has a higher change point value and higher mortality. The proposed model, CP-SISM, with one common change point for all cities, is compared to the aggregated data model that is commonly used in previous studies, and the proposed model was found to be much better in terms of fitting the data (higher R^2) and detecting the significant spatial change point (smaller p-value and standard error). The proposed model, CP-SISM, with possible spatially dependent change points, is compared to the case of each city's data separately analyzed to detect its change points, which is considered in many previous studies. It is found that the change-point values are comparable, but three cities have insignificant change points in the case of separately analyzed city data (previous studies have not tested the change points detected) and the change-point estimates of the CP-SISM have smaller standard errors and smaller p-values.

The proposed model with one common change point is compared to the spatially dependent change points case. Both models showed that Busan City has the highest mortality, and Seoul and Daejeon have the lowest mortality. The CP-SISM with spatially-dependent change points has a higher R^2 value and detected that one of the cities (i.e., Busan) has a higher change point compared to the other cities.

The proposed model offers several opportunities for extension and enhancements to improve the estimation method. The proposed model assumes the mean mortality function over the cities has the same shape; however, this assumption can be relaxed and can use different functions for different cities.

In the proposed model, it is assumed that the mean mortality functions, $f(\cdot)$, over cities, have the same form and then detect and test the change points. It is possible to consider change point detection for the non-parametric part $f(\cdot)$. This approach would involve identifying shifts in the functional form or underlying structure of $f(\cdot)$ over locations. Implementing change point detection in a non-parametric context, however, may require different techniques, or other non-parametric hypothesis tests, to effectively capture and detect changes in $f(\cdot)$. The model assumes that the slopes before and after the change point are the same for all cities, but different slopes can be used. In the model estimation, a grid search is used to obtain the change points, however, better methods can be used such as assuming the change points are random variables following some distribution with some mean and variance, such as a normal distribution. This will reduce the estimation time, especially if the Bayesian approach is used. The proposed approach is applied to 6 cities in South Korea, but it can be applied to cities from different countries. In some countries, the spatial effects may be integrated into the mean function as follows:

$$\boldsymbol{\mu}|\mathbf{u},\boldsymbol{\beta},\theta \quad = \quad f(X(\boldsymbol{\theta})\boldsymbol{\beta} + Z\mathbf{u}).$$

In this case, there will be no identifiability problem for the single index function, and for some countries, the spatial random effects may not be additive to the nonparametric function. Mortality was found to depend on pollutant and weather variables as an index (a linear combination of these variables). In this context, a variable selection method can identify significant index variables affecting mortality. To address the identifiability issue and facilitate variable selection, the constraint $||\beta|| = 1$ can be applied instead of fixing the first parameter of β to 1. The proposed model can be extended to accommodate generalized linear models beyond the Poisson framework. For instance, when the response variable is binary, methods designed for estimating single-index functions can be applied using Bernoulli distribution. Once the single-index model is estimated using such an approach, the subsequent steps in the proposed methodology become straightforward.

Environmental epidemiology often provides high-dimensional variables so we need to detect many change points. In this case, we can build a high-dimensional nonparametric model using deep neural network tools and visualize these high-dimensional change points using computer vision. These connections among machine learning architecture [37], computer vision [38, 39], and statistical models will provide more flexible analytical tools for complex data.

Supporting information

S1 Table. Characteristics of the 6 major cities in Korea: Seoul, Busan, Daegu, Incheon, Gwangju, and Daejeon.
(PDF)

Author Contributions

Conceptualization: Inyoung Kim.

Data curation: Inyoung Kim.

Formal analysis: Hamdy F. F. Mahmoud.

Methodology: Hamdy F. F. Mahmoud, Inyoung Kim.

Software: Hamdy F. F. Mahmoud.

Supervision: Inyoung Kim.

Validation: Hamdy F. F. Mahmoud, Inyoung Kim.

Visualization: Hamdy F. F. Mahmoud.

Writing – original draft: Hamdy F. F. Mahmoud.

Writing – review & editing: Hamdy F. F. Mahmoud, Inyoung Kim.

References

1. Anderson BG, Bell ML. Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States. Epidemiology. 2009; 20: 205–213. https://doi.org/10.1097/EDE.0b013e318190ee08 PMID: 19194300
2. Wang C, Chen R, Kuang X, Duan X, Kan H. Temperature and daily mortality in Suzhou, China: a time series analysis. Sci Total Environ. 2014 Jan; 466-467: 985–990. https://doi.org/10.1016/j.scitotenv.2013.08.011 PMID: 23994732
3. Gasparrini A, Guo Y, Hashizume M, Lavigne E, Zanobetti A, Schwartz J, et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. Lancet. 2015; 386:369–375. https://doi.org/10.1016/S0140-6736(14)62114-0 PMID: 26003380
4. Basu R, and Samet JM. Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence. Epidemiology Review. 2002; 24:190–202. https://doi.org/10.1093/epirev/mxf007 PMID: 12762092
5. Armstrong B. Models of the relationship between ambient temperature and daily mortality. Epidemiology. 2006; 17:624–631. https://doi.org/10.1097/01.ede.0000239732.50999.8f PMID: 17028505
6. Hashizume M, Wagatsuma Y, Hayashi T, Saha S, Streatfield K, Yunus M. The effect of temperature on mortality in rural Bangladesh a population-based time series study. International Journal of Epidemiology. 2009; 38:1689–1697. https://doi.org/10.1093/ije/dyn376 PMID: 19181749
7. Son J, Lee J, and Anderson GB, Bell ML. Vulnerability to temperature-related mortality in Seoul, Korea. Environmental Research Letters. 2011; 6:1–8. https://doi.org/10.1088/1748-9326/6/3/034027 PMID: 23335945
8. El-Zein A, Tewtel-Salem M, Nehme G. A time-series analysis of mortality and air temperature in Greater Beirut. Science of the Total Environment. 2004; 330:71–80. https://doi.org/10.1016/j.scitotenv.2004.02.027 PMID: 15325159

9. Hattis D, Ogneva-Himmelberger Y, Ratick S. The spatial variability of heat-related mortality in Massachusetts. Applied Geography. 2012; 33: 45–52. https://doi.org/10.1016/j.apgeog.2011.07.008
10. Kan H, London SJ, Chen H, Song G, Chen G, Jiang L, et al. Diurnal temperature range and daily mortality in Shanghai, China. Environmental Research. 2007; 103: 424–431. https://doi.org/10.1016/j.envres.2006.11.009 PMID: 17234178
11. Mahmoud HFF, Kim I, Kim H. Semiparametric single index multi change points model with an application of environmental health study on mortality and temperature. Environmetrics. 2016; 27(8):494–506. https://doi.org/10.1002/env.2413
12. Nakai S, Itoh T, Morimoto T. Deaths from heat-stroke in Japan: 1968-1994. International Journal of Biometeorol. 1999; 43: 124–127. https://doi.org/10.1007/s004840050127 PMID: 10639904
13. Hajat S, Kovats RS, Atkinson RW, Heines A. Impact of hot temperatures on death in London: a time series approach. Journal Epidemiology Community Health. 2002; 56: 367–372. https://doi.org/10.1136/jech.56.5.367
14. Curriero FC, Heiner KS, Samet JM, Zeger SL, Strug L, Patz JA. Temperature and mortality in 11 cities of the eastern United States. American Journal of Epidemiology. 2002; 155:80–87. https://doi.org/10.1093/aje/155.1.80 PMID: 11772788
15. Kim H, Ha J-S, Park J. High Temperature, Heat Index, and Mortality in 6 Major Cities in South Korea. Environmental and Occupational Health. 2006; 61(6):265. https://doi.org/10.3200/AEOH.61.6.265-270 PMID: 17967749
16. Chung J, Honda Y, Hong Y, Pan X, Guo Y, Kim H. Ambient temperature and mortality: an international study in four capital cities of East Asia. Science of the Total Environment. 2009; 408:390–396. https://doi.org/10.1016/j.scitotenv.2009.09.009 PMID: 19853280
17. Baccini M, Biggeri A, Accetta G, Kosatsky T, Katsouyanni K, Analitis A, et al. Heat Effects on Mortality in 15 European Cities. Epidemiology. 2008; 19(5):711–719. https://doi.org/10.1097/EDE.0b013e318176bfcd PMID: 18520615
18. Lee W-H, Lim Y-H, Dang TN, Seposo x, Honda Y, Guo Y-LL, et al. An Investigation on Attributes of Ambient Temperature and Diurnal Temperature Range on Mortality in Five East-Asian Countries. Scientific Reports. 2017; 7(1):1–9. https://doi.org/10.1038/s41598-017-10433-8 PMID: 28860544
19. MA C, Honda Y, Dang TN. Comparison of wet-bulb globe temperature (WBGT) and mean temperature for assessment of heat-related mortality: evidence from 47 Japanese prefectures. Japanese Journal of Health and Human Human Ecology. 2018; 84(2): 52–72. https://doi.org/10.3861/kenko.84.2_52
20. Luan G, Yin P, Wang L, Zhou M. The temperature-mortality relationship: an analysis from 31 Chinese provincial capital cities. International Journal of Environmental Health Research. 2018; 28(2). https://doi.org/10.1080/09603123.2018.1453056 PMID: 29562755
21. Cressie, N. Statistics for Spatial Data. Wiley, 1993.
22. Banerjee, S, Carlin, BP, Gelfand, AE. HHierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC. 2004.
23. Diggle PJ, and Ribeiro PJ. Model-Based Geostatistics. Springer, 2007.
24. Gelfand AE, Diggle P, Fuentes M, Guttorp P. Mandbook of Spatial Statistics. CRC Press, 2010.
25. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics. 1999; 58:71–120. https://doi.org/10.1016/0304-4076(93)90114-K
26. Sherman RP. U-process in analysis of a generalized semi-parametric regression estimator. Economic theory. 1994; 10:372–395. https://doi.org/10.1017/S0266466600008458
27. Lin W, Kulasekera KB. Identifiability of single index models and additive index models. Biometrika. 2007; 94: 496–501. https://doi.org/10.1093/biomet/asm029
28. Xia Y, Li WK, Tong H, Zhang D. A goodness-of-fit for single index models. Statistica sinica. 2004; 14:1–39.
29. Hardle W, Hall P, Ichimura H. Optimal smoothing in single index models. The analysis of statistics. 1993; 21: 157–178.
30. McCulloch CE. Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association. 1994; 89:330–335. https://doi.org/10.1080/01621459.1994.10476474
31. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association. 1997; 92:162–170. https://doi.org/10.1080/01621459.1997.10473613
32. Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo em algorithm. Journal of the Royal Statistical Society-B. 1999; 61: 265–285. https://doi.org/10.1111/1467-9868.00176

33. Caffo BS, Jank W, Jones GL. Ascent-based Monte Carlo expectation maximization. Journal of the Royal Statistical Society, Series B. 2005; 67: 235–251. https://doi.org/10.1111/j.1467-9868.2005.00499.x

34. Tan M, Tian G-L, Fang H-B. An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data. Journal of Statistical Computation and Simulation. 2007; 77: 929–943. https://doi.org/10.1080/10629360600843153

35. An X, Bentler PM. Efficient direct sampling MCEM algorithm for latent variable models with binary responses. Computational Statistics and Data Analysis. 2012; 56: 231–244. https://doi.org/10.1016/j.csda.2011.06.028

36. Murage P, Hajat S, Bone A. Variation in Cold-Related Mortality in England Since the Introduction of the Cold Weather Plan: Which Areas Have the Greatest Unmet Needs?. International Journal of Environmental Research and Public Heath. 2018; 15(11): 2528. https://doi.org/10.3390/ijerph15112588 PMID: 30463273

37. Zhang J, Su Q, Tang B, Wang C, Li Y. DPSNet: Multitask Learning Using Geometry Reasoning for Scene Depth and Semantics. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS. 2021; 3107362

38. Liu Y, Zhang J. Service Function Chain Embedding Meets Machine Learning: Deep Reinforcement Learning Approach. IEEE Transactions on Network and Service Management, 2024. https://doi.org/10.1109/TNSM.2024.3353808

39. Zhang J, Huang S, Liu J, Zhu X, Xu F. PYRF-PCR: A Robust Three-Stage 3D Point Cloud Registration for Outdoor Scene. IEEE Transactions on Intelligent Vehicles. https://doi.org/10.1109/TIV.2023.3327098