A Novel Spatial Interpolation Method Based on Spatio-Temporal Data and Cointegration Theory Along with an R Package TSCS

by TIANJIAN YANG

June 30, 2017

Abstract

We propose an innovative method for spatial interpolation called TSCS (abbr. of Time Series Cointegrated System), which is based on cointegration theory and multiple linear regression. It considers long-term equilibrium relationship and requires making use of historical spatio-temporal data, though it is a purely spatial interpolation method. TSCS bears two main advantages. Firstly, it generally performs well when making interpolation and possesses high robustness. Furthermore, it is relatively simple and easy to implement without model selection, parameter adjustment or requirement of subjective judgement, giving it a chance to be regarded as a desirable alternative to existing spatio-temporal interpolation methods in some cases where we merely intend to interpolate a series of cross-section data at each observed time point for a given spatial domain.

The theory framework of TSCS is presented first. Then, through simulation study, we show its high accuracy of interpolation along with good robustness. Furthermore, some properties of TSCS referring to its performance are studied via repeated experiments. Next, TSCS is compared with spatio-temporal kriging in a real-world application, based on the GHCND data set, concerning spatio-temporal interpolation, which illustrates the prominent strengths of TSCS in some specific cases.

Additionally, an R package named **TSCS** is built for carrying out TSCS spatial interpolation method. R code of each function is presented in simulation study to demonstrate the workflow of TSCS using this package.

Key words: spatial statistics, geostatistics, spatio-temporal data, time series, regression

1 Introduction

The collection and processing of spatio-temporal data is rapidly increasing due to technological advances and the societal need for analysis of variables that vary in space and time, such as weather and air quality variables. Nowadays, modern sensors allow to monitor different variables at an increasing temporal resolution producing rich spatio-temporal data sets. With ubiquitous spatio-temporal data, relevant problems about how to comprehend and make full use of them have also penetrated into various fields. Purely spatial interpolation (Hua Xu 2012), like kriging (Stein, M. L 1999), if resorting to spatio-temporal data can potentially provide more accurate predictions than without considering them because observations taken at other times can be included for analysis.

Bigger data set leads to more time-consuming algorithm and, meanwhile, taking time into account makes model much more complex. It is fairly difficult to find the best or the most convenient model when dealing with spatio-temporal data. Accordingly, with the recent development of geostatistics, meteorology and econometrics, many methods have been

proposed to handle spatio-temporal data, struggling for spatio-temporal interpolation, such as spatio-temporal kriging (Benedikt Graler et al. 2016), STARMA (Felix Cheysson 2016), random field (Martin Schlather et al. 2015) and Gaussian spatio-temporal process (Johan Lindstrom et al. 2013). They are widely accepted and all perform well on a large class of problems respectively. However, after applying the above methods to real data, we discern two main problems. One is that these methods are very complicated based on various intricate models and substantial parameters. They more or less require subjective judgements from human and a cumbersome process of model selection. The other problem is that these methods fail to efficaciously deal with a class of spatio-temporal interpolation problem, where the historical spatio-temporal data is relatively full while the new spatio-temporal data, with numerous missing observations that we want to interpolate at each observed time point, is a series of cross-section data which is sparse in time dimension (see Fig 1). Since the new spatio-temporal data to be interpolated is sparse in time dimension, it is hard to establish a model of good fit analyzing historical data and new data simultaneously. Hence, the performance of interpolation would not be good enough. Moreover, if the time interval between historical data and new data is large, the performance will only get worse.

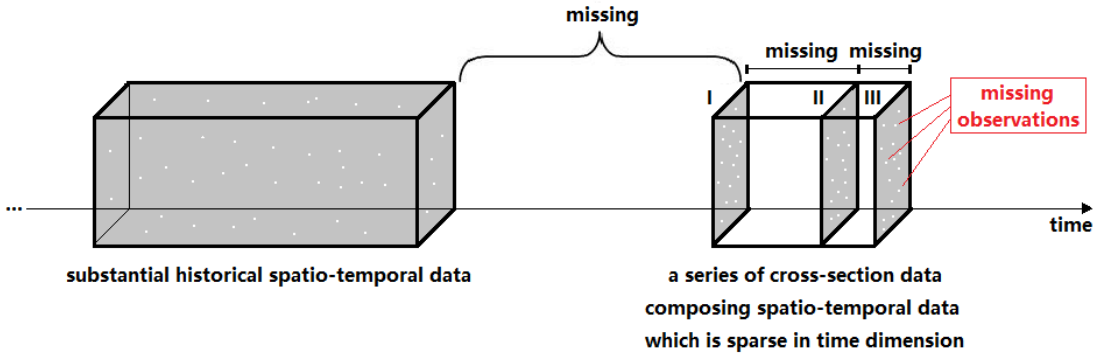

Fig 1. Problem clarification.

As a consequence, we propose a new method called TSCS (abbr. of Time Series Cointegrated System) to cope with the problems discussed above. This method possesses two advantages. On the one hand, it generally performs well when making interpolation and possesses high robustness. On the other hand, it is simple and easy to employ without the need of model selection, parameter adjustment and subjective judgement.

Theoretically, TSCS derives from cointegration theory and multiple linear regression. In the process of coming up with this method, we shift our perspective towards cointegration and spatio-temporal data simultaneously in an outside-the-box manner, a far cry from existing spatial interpolation methods and spatio-temporal modelling theory. The main consideration of TSCS is the long-term equilibrium relationship between spatial locations, instead of involving spatio-temporal covariance function or covariate in most spatio-temporal modelling methods (Johan Lindstrom et al. 2013; Martin Schlather et al. 2015), variogram and anisotropy in spatio-temporal kriging (Benedikt Graler et al. 2016), etc.

The rest of this paper is organized as follows. In Section 2, we describe the core idea of TSCS method first. Then, we make clear the definitions of key concepts. Next, the detailed process of TSCS and its theory are demonstrated. In Section 3, based on specially simulated data, we show the result of simulation study from 2D and 3D rectangular grid system separately. Meanwhile, the workflow of spatial interpolation using R package **TSCS** is presented. In Section 4, we compare TSCS with spatio-temporal kriging, one of the state-of-the-art spatio-temporal interpolation methods, in a real-world application concerning the GHCND data set, to illustrate the strengths of TSCS. The paper is concluded by some remarks in Section 5, where the drawbacks of TSCS and some unsettled problems are tersely stated.

2 Method

2.1 Overview of TSCS

In the context of geostatistics (Donald E. Myers), a natural spatio-temporal data set can be an observation set yielded from space geodetic system, meteorological observing system or just a simple farmland monitor net. Under the circumstance we mentioned before, suppose the spatio-temporal data set to be interpolated is composed of a series of cross-section data which is sparse in time dimension, with numerous missing observations that we intend to predict at each observed time point. Meanwhile, we have enough historical spatio-temporal data in hand. (see Fig 1)

TSCS is based on cointegration theory and multiple linear regression, the core idea of which is as follows. In the spatial domain of the data set mentioned above, we consider that every spatial location includes an individual time series. Theoretically, we first assume that, for any spatial location within spatial domain, its time series and the time series of its adjacent spatial locations are cointegrated (cointegrated system). Then, based on historical spatio-temporal data, for each spatial location along with its adjacent locations, we calculate the regression coefficients through multiple linear regression. Finally, with the use of the regressive relationship obtained, through establishing system of linear equations and solving it, missing observations are estimated. Our reasoning is that the regression function obtained on the strength of historical spatio-temporal data is able to explicate the long-term equilibrium relationships between spatial locations, which means the correlations still hold in the future if the system property is relatively stable. Thus, we can utilize the regression coefficients for missing value prediction.

Strictly speaking, TSCS is not a general spatio-temporal interpolation method for two reasons. First, as we have emphasized before, the actual work of TSCS is making interpolation in a new data set (posterior) on the basis of analyzing historical spatio-temporal data (anterior) in hand (see Fig 1), rather than purely interpolating within a single data set. Second, in the process of estimating missing observations, TSCS handles each cross-section data separately, which is unable to give prediction in a time point whose cross-section data is entirely absent. Hence, it should be regarded as a purely spatial interpolation method but includes the consideration of time series and cointegrated relationship.

2.2 Definitions of Key Concepts

The following key concepts will be used repeatedly in this paper. ***Definition 3*** are created in the context of TSCS exclusively, for the convenience of further statement. The others are existing in academia.

Definition 1. Stationary (weak stationary) time series. A stationary time series x_t is a finite variance process such that:

(i) the mean value function μ_t is constant and does not depend on time t.

(ii) the autocovariance function $\gamma(s,t)$ depends on time s and t only through their difference $|s-t|$.

Definition 2. Cointegrated relationship. Cointegration is a statistical property of a collection $(x_{1t}, x_{2t}, \ldots, x_{kt})$ of time series variables. First, all of the series must be integrated of order one. Next, if there exists a linear combination of this collection integrated of order zero, then the collection is said to be cointegrated. Cointegration means long-term equilibrium relationship.

Definition 3. Cointegrated system. In the context of TSCS, cointegrated system is a class of spatio-temporal data. Firstly, in the spatial domain of the data, we consider that every spatial location includes an individual time series (missing values are allowed). If the data satisfies that, for any spatial location, its time series and the time series of its adjacent spatial locations are cointegrated, it is said to be a cointegrated system.

Definition 4. Cross-section data. Cross-section data is a type of data collected by observing many subjects (for example, spatial locations in the context of geostatistics) at the same point of time, without regard to differences in time. (see Fig 2)

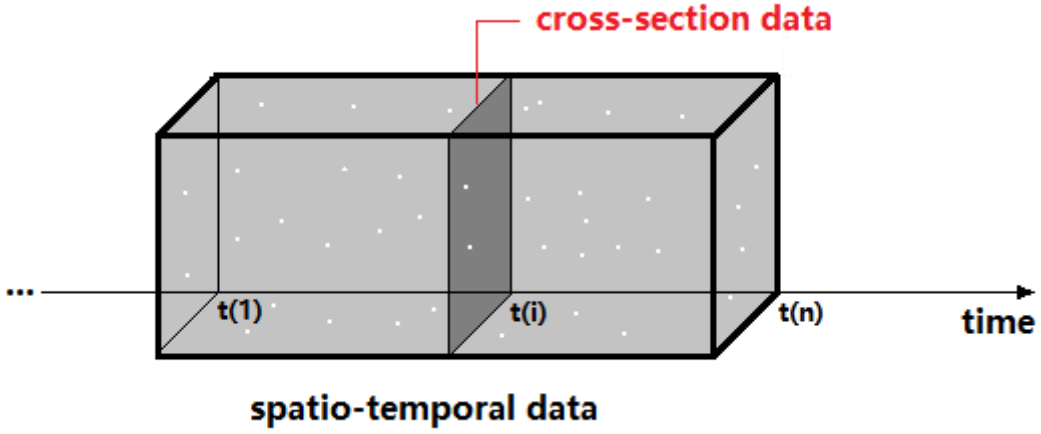

Fig 2. Definition illustration.

2.3 Detailed Statement of TSCS

In the following *2.3.1 ~ 2.3.4*, we clearly elaborate the four essential steps of TSCS procedure in order, but we decide to take 2D rectangular grid system for example only. This is because the theory and procedure stays identical no matter for 2D or 3D rectangular grid system and the only difference is the selection of adjacent spatial locations along with algorithmic details. We don't need to make redundant pages.

Suppose, in a given 2D rectangular grid system, we have the following spatio-temporal data.

(i) Spatial domain $\mathbf{S}$ along with coordinates of spatial locations: (X_i, Y_i), $i = 1, 2, \ldots, n$

(ii) Spatio-temporal data in temporal domain $\mathbf{T}$: $\mathbf{Z} = \{z(s_1, t_1), z(s_2, t_2), \ldots, z(s_m, t_m)\}$, where $(s_1, t_1), (s_2, t_2), \ldots, (s_m, t_m) \in \mathbf{S} \times \mathbf{T} \subseteq \mathbb{R}^2 \times \mathbb{R}$

(iii) Spatio-temporal data in temporal domain $\mathbf{T}^*$: $\mathbf{Z}^* = \{z(s_1, t_1^*), z(s_2, t_2^*), \ldots, z(s_p, t_p^*)\}$, where $(s_1, t_1^*), (s_2, t_2^*), \ldots, (s_m, t_m^*) \in \mathbf{S} \times \mathbf{T}^* \subseteq \mathbb{R}^2 \times \mathbb{R}$

In addition, $\mathbf{T}$ is anterior to $\mathbf{T}^*$ in time axis. Consequently, $\mathbf{Z}$ is called historical spatio-temporal data while $\mathbf{Z}^*$ is the new spatio-temporal data we are interested in, with missing observations to be interpolated.

2.3.1 Cointegration Test

The basic assumption of TSCS method is that both $\mathbf{Z}$ and $\mathbf{Z}^*$ enjoying the same spatial domain $\mathbf{S}$ can be regarded as a cointegrated system. Hence, cointegration test is necessary before TSCS is put into use.

Spurious relationship (Anindya Banerjee et al. 1993) is a common problem in statistics. As is known to all, whether a valid regression model can be established between a collection of time series depends on if cointegrated relationship exists between them. Provided that these time series are not cointegrated, the residual of fitting is a nonstationary time series leading to spurious relationship. In this case, the regression function obtained cannot truly explain the long-term equilibrium relationship between these variables even if it is a good fit statistically. Therefore, cointegration test must be done before building regression model between multiple time series (Shiying Zhang et al. 2014).

The prerequisite of cointegration test is that all of the time series considered must be integrated of one, namely, first-order difference stationary. As a result, after setting significance level, we first test first-order difference stationarity of time series for every spatial location in spatial domain. Next, within the whole spatial domain, only if for any spatial location, its time series bears a cointegrated relationship with the time series of its adjacent spatial locations, proved by cointegration test, can it make sense to use the regression functions obtained for estimation later.

Detailed procedure about cointegration test is demonstrated in the next part *2.3.2* in that regression analysis is in correspondence with estimating cointegration coefficients by means of OLS (Orthogonal Least Square) method (Shiying Zhang et al. 2014).

2.3.2 Obtaining Regression Coefficients Matrix

To begin with, we assert that TSCS is only capable of interpolation but not extrapolation. TSCS is unable to estimate missing observation located in the boundary or beyond the range of a given spatial domain.

Since there is no distinction between obtaining regression coefficients and calculating cointegration coefficients through OLS method, after distinguishing interior spatial locations from spatial domain boundary, we establish the following regression model (Michael H. Kutner et al. 1988) for every interior spatial location $s_i \in \mathbf{S}$ along with its J adjacent spatial locations $s_{i(1)}, s_{i(2)}, \ldots, s_{i(J)} \in \mathbf{S}$.

$$z(s_i, t) = \beta_0 + \sum_{j=1}^{J} \beta_j z(s_{i(j)}, t) + \varepsilon_t$$

In this model, ε_t denotes random error term that satisfies $E(\varepsilon_t) = 0$ and $\mathrm{var}(\varepsilon_t) = \sigma^2 < \infty$, ε_u and ε_v are uncorrelated $(u \neq v)$. Since we are taking a 2D rectangular grid system as an exemplificative case, the selection of adjacent spatial locations can be decided as what Fig 3 shows. Thus, here $J = 8$.

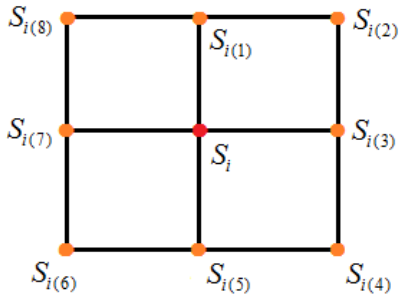

Fig 3. Selection of adjacent spatial locations for 2D rectangular grid system.

Based on sample $z(s_i, t), z(s_{i(1)}, t), z(s_{i(2)}, t), \ldots, z(s_{i(8)}, t)$, $t \in \mathbf{T}$ from historical spatio-temporal data $\mathbf{Z}$, we obtain the following fitting function.

$$z(s_i,t)=\hat{\beta}_0+\sum_{j=1}^{8}\hat{\beta}_j z(s_{i(j)},t)+e_t$$

According to cointegration theory, the behavior of residual e_t determines whether spurious correlation occurs. For a given significance level generally set to 0.05, through unit root test (Alok Bhargava 1986), if we conclude that e_t is stationary, we can say the collection of time series $z(s_i,t), z(s_{i(1)},t), z(s_{i(2)},t),\ldots, z(s_{i(8)},t)$, $t\in\mathbf{T}$ is cointegrated. Under cointegrated system (see ***Definition 3***), the basic assumption of TSCS method, using historical spatio-temporal data $\mathbf{Z}$, we obtain the regression coefficient vector $\boldsymbol{B}=(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_8)$ for every interior spatial location, stored into a matrix $\boldsymbol{\Theta}$. The regression coefficient matrix $\boldsymbol{\Theta}$ is used for missing value estimation in spatio-temporal data $\mathbf{Z}^*$.

2.3.3 Searching for Subdomain with Missing Observation

In the process of estimating missing observation in spatio-temporal data $\mathbf{Z}^*$, TSCS works on each cross-section data separately and it only deals with the subdomain with missing observation, without involving other parts of the spatial domain. Hence, we need to search out these subdomains with missing observation inside each cross-section data in $\mathbf{Z}^*$.

For the sake of simplicity, we might as well focus on one cross-section data $\mathbf{Z}^*(t_h^*)\subseteq\mathbf{Z}^*$ at time of $t_h^*\in\mathbf{T}^*$ as an example (see Fig 4), where the missing observations are denoted by following notations.

$$y_1=z(s_{h(1)},t_h^*),\ y_2=z(s_{h(2)},t_h^*),\ldots,\ y_9=z(s_{h(9)},t_h^*)\notin\mathbf{Z}^*(t_h^*)$$
$$s_{h(1)},s_{h(2)},\ldots,s_{h(9)}\in\mathbf{S}\ \text{ and }\ t_h^*\in\mathbf{T}^*$$

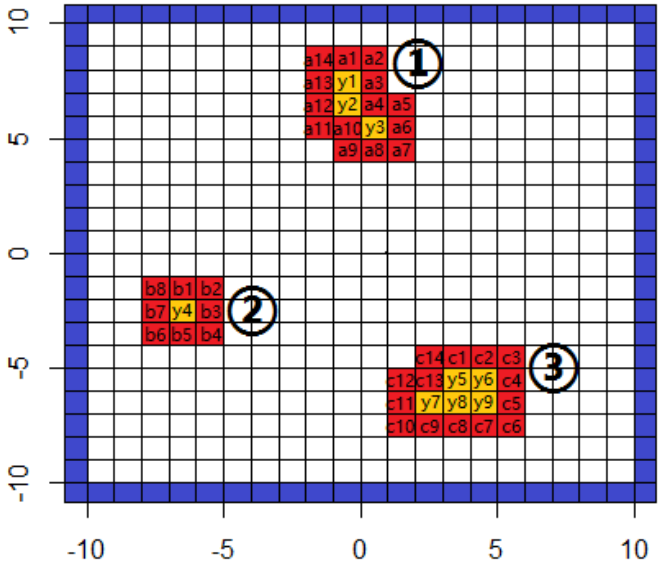

Fig 4. Three subdomains with missing observation in one cross-section data.
Each lattice denotes a spatial location. Blue lattice refers to spatial domain boundary.
Red lattice refers to subdomain boundary. Yellow lattice designates missing observation.

According to the spatial distribution of $y_1, y_2,\ldots, y_9$, we discover three subdomains with missing observation in total, denoted by $\mathbf{S}_1=\{s_{h(1)},s_{h(2)},s_{h(3)}\}$, $\mathbf{S}_2=\{s_{h(4)}\}$ and

$\mathbf{S}_3=\{s_{h(5)}, s_{h(6)}, s_{h(7)}, s_{h(8)}, s_{h(9)}\}$, where $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3 \subseteq \mathbf{S}$. Here it is quite necessary to emphasize that the missing observation can be spatially isolated or contiguous in group within subdomain.

2.3.4 Solving System of Linear Equations

Now we proceed to the last stage because all subdomains with missing observation have been searched out. With the use of regression coefficient matrix $\boldsymbol{\Theta}$ obtained from historical spatio-temporal data $\mathbf{Z}$, in combination with the cross-section data at time of t_h^*, we establish the following system of linear equations for each subdomain $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$, where $y_1, y_2, \ldots, y_9$ are considered as variables. The meanings of $a_1 \sim a_{14}$, $b_1 \sim b_8$ and $c_1 \sim c_{14}$ are demonstrated in Fig 4.

$$\Phi_1: \begin{cases} y_1 = \hat{\beta}_{01} + \hat{\beta}_{11}a_1 + \hat{\beta}_{21}a_2 + \hat{\beta}_{31}a_3 + \hat{\beta}_{41}a_4 + \hat{\beta}_{51}y_2 + \hat{\beta}_{61}a_{12} + \hat{\beta}_{71}a_{13} + \hat{\beta}_{81}a_{14} \\ y_2 = \hat{\beta}_{02} + \hat{\beta}_{12}y_1 + \hat{\beta}_{22}a_3 + \hat{\beta}_{32}a_4 + \hat{\beta}_{42}y_3 + \hat{\beta}_{52}a_{10} + \hat{\beta}_{62}a_{11} + \hat{\beta}_{72}a_{12} + \hat{\beta}_{82}a_{13} \\ y_3 = \hat{\beta}_{03} + \hat{\beta}_{13}a_4 + \hat{\beta}_{23}a_5 + \hat{\beta}_{33}a_6 + \hat{\beta}_{43}a_7 + \hat{\beta}_{53}a_8 + \hat{\beta}_{63}a_9 + \hat{\beta}_{73}a_{10} + \hat{\beta}_{83}y_2 \end{cases}$$

$$\Phi_2: \; y_4 = \hat{\beta}_{04} + \hat{\beta}_{14}b_1 + \hat{\beta}_{24}b_2 + \hat{\beta}_{34}b_3 + \hat{\beta}_{44}b_4 + \hat{\beta}_{54}b_5 + \hat{\beta}_{64}b_6 + \hat{\beta}_{74}b_7 + \hat{\beta}_{84}b_8$$

$$\Phi_3: \begin{cases} y_5 = \hat{\beta}_{05} + \hat{\beta}_{15}c_1 + \hat{\beta}_{25}c_2 + \hat{\beta}_{35}y_6 + \hat{\beta}_{45}y_9 + \hat{\beta}_{55}y_8 + \hat{\beta}_{65}y_7 + \hat{\beta}_{75}c_{13} + \hat{\beta}_{85}c_{14} \\ y_6 = \hat{\beta}_{06} + \hat{\beta}_{16}c_2 + \hat{\beta}_{26}c_3 + \hat{\beta}_{36}c_4 + \hat{\beta}_{46}c_5 + \hat{\beta}_{56}y_9 + \hat{\beta}_{66}y_8 + \hat{\beta}_{76}y_5 + \hat{\beta}_{86}c_1 \\ y_7 = \hat{\beta}_{07} + \hat{\beta}_{17}c_{13} + \hat{\beta}_{27}y_5 + \hat{\beta}_{37}y_8 + \hat{\beta}_{47}c_8 + \hat{\beta}_{57}c_9 + \hat{\beta}_{67}c_{10} + \hat{\beta}_{77}c_{11} + \hat{\beta}_{87}c_{12} \\ y_8 = \hat{\beta}_{08} + \hat{\beta}_{18}y_5 + \hat{\beta}_{28}y_6 + \hat{\beta}_{38}y_9 + \hat{\beta}_{48}c_7 + \hat{\beta}_{58}c_8 + \hat{\beta}_{68}c_9 + \hat{\beta}_{78}y_7 + \hat{\beta}_{88}c_{13} \\ y_9 = \hat{\beta}_{09} + \hat{\beta}_{19}y_6 + \hat{\beta}_{29}c_4 + \hat{\beta}_{39}c_5 + \hat{\beta}_{49}c_6 + \hat{\beta}_{59}c_7 + \hat{\beta}_{69}c_8 + \hat{\beta}_{79}y_8 + \hat{\beta}_{89}y_5 \end{cases}$$

Next, we combine the three systems of linear equations into one joint system of linear equations. By transposition, its characteristics are more explicit.

$$\text{Let} \begin{cases} \hat{\beta}_{01} + \hat{\beta}_{11}a_1 + \hat{\beta}_{21}a_2 + \hat{\beta}_{31}a_3 + \hat{\beta}_{41}a_4 + \hat{\beta}_{61}a_{12} + \hat{\beta}_{71}a_{13} + \hat{\beta}_{81}a_{14} = I_1 \\ \hat{\beta}_{02} + \hat{\beta}_{22}a_3 + \hat{\beta}_{32}a_4 + \hat{\beta}_{52}a_{10} + \hat{\beta}_{62}a_{11} + \hat{\beta}_{72}a_{12} + \hat{\beta}_{82}a_{13} = I_2 \\ \hat{\beta}_{03} + \hat{\beta}_{13}a_4 + \hat{\beta}_{23}a_5 + \hat{\beta}_{33}a_6 + \hat{\beta}_{43}a_7 + \hat{\beta}_{53}a_8 + \hat{\beta}_{63}a_9 + \hat{\beta}_{73}a_{10} = I_3 \\ \hat{\beta}_{04} + \hat{\beta}_{14}b_1 + \hat{\beta}_{24}b_2 + \hat{\beta}_{34}b_3 + \hat{\beta}_{44}b_4 + \hat{\beta}_{54}b_5 + \hat{\beta}_{64}b_6 + \hat{\beta}_{74}b_7 + \hat{\beta}_{84}b_8 = I_4 \\ \hat{\beta}_{05} + \hat{\beta}_{15}c_1 + \hat{\beta}_{25}c_2 + \hat{\beta}_{75}c_{13} + \hat{\beta}_{85}c_{14} = I_5 \\ \hat{\beta}_{06} + \hat{\beta}_{16}c_2 + \hat{\beta}_{26}c_3 + \hat{\beta}_{36}c_4 + \hat{\beta}_{46}c_5 + \hat{\beta}_{86}c_1 = I_6 \\ \hat{\beta}_{07} + \hat{\beta}_{17}c_{13} + \hat{\beta}_{47}c_8 + \hat{\beta}_{57}c_9 + \hat{\beta}_{67}c_{10} + \hat{\beta}_{77}c_{11} + \hat{\beta}_{87}c_{12} = I_7 \\ \hat{\beta}_{08} + \hat{\beta}_{48}c_7 + \hat{\beta}_{58}c_8 + \hat{\beta}_{68}c_9 + \hat{\beta}_{88}c_{13} = I_8 \\ \hat{\beta}_{09} + \hat{\beta}_{29}c_4 + \hat{\beta}_{39}c_5 + \hat{\beta}_{49}c_6 + \hat{\beta}_{59}c_7 + \hat{\beta}_{69}c_8 = I_9 \end{cases}$$

$$\text{Then} \begin{cases} \Phi_1 \\ \Phi_2 \\ \Phi_3 \end{cases} = \begin{cases} y_1 - \hat{\beta}_{51} y_2 = I_1 \\ -\hat{\beta}_{12} y_1 + y_2 - \hat{\beta}_{42} y_3 = I_2 \\ -\hat{\beta}_{83} y_2 + y_3 = I_3 \\ y_4 = I_4 \\ y_5 - \hat{\beta}_{35} y_6 - \hat{\beta}_{65} y_7 - \hat{\beta}_{55} y_8 - \hat{\beta}_{45} y_9 = I_5 \\ -\hat{\beta}_{76} y_5 + y_6 - \hat{\beta}_{66} y_8 - \hat{\beta}_{56} y_9 = I_6 \\ -\hat{\beta}_{27} y_5 + y_7 - \hat{\beta}_{37} y_8 = I_7 \\ -\hat{\beta}_{18} y_5 - \hat{\beta}_{28} y_6 - \hat{\beta}_{78} y_7 + y_8 - \hat{\beta}_{38} y_9 = I_8 \\ -\hat{\beta}_{89} y_5 - \hat{\beta}_{19} y_6 - \hat{\beta}_{79} y_8 + y_9 = I_9 \end{cases}$$

The solution of the above joint system of linear equations is the estimation of missing observations $y_1, y_2, \ldots, y_9$. Similarly, for every cross-section data in $\mathbf{Z}^*$ to be interpolated, we estimate its missing observations in the way demonstrated above. Up to now, TSCS interpolation is completed. We can see that TSCS is a spatial interpolation method in nature, but it is able to solve part of spatio-temporal interpolation problem.

On balance, under cointegrated system $\mathbf{Z}$ and $\mathbf{Z}^*$, we obtain regression coefficient matrix $\boldsymbol{\Theta}$ based on $\mathbf{Z}$ first. Suppose the cross-section data $\mathbf{Z}^*(t_h^*)$ at time of t_h^* has missing observations $y_1, y_2, \ldots, y_K$. Then, we search out subdomains with missing observation $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{K'}$, $K' \le K$. Next, using $\boldsymbol{\Theta}$ and $\mathbf{Z}^*(t_h^*)$, we construct system of linear equations $\Phi_1, \Phi_2, \ldots, \Phi_{K'}$ with regard to subdomains $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{K'}$. Finally, by solving the joint system of linear equations $\{\Phi_1, \Phi_2, \ldots, \Phi_{K'}\}$, we obtain the estimation of missing observations $y_1, y_2, \ldots, y_K$. In this way, we give prediction to missing observation in every cross-section data of $\mathbf{Z}^*$. Hereto, interpolation by means of TSCS is totally completed.

3 Simulation Study

The three main objectives of simulation study are as follows.

First, in **3.2**, to demonstrate the workflow of spatial interpolation using **TSCS**, an R package tailored for TSCS method. Here we emphasize that **TSCS** of current version 0.1.1 is only capable of handling spatio-temporal data based on 2D and 3D rectangular grid system, two typical cases common in real life. This package can be downloaded from CRAN repository at URL http://CRAN.R-project.org/package=TSCS. Additionally, other R packages that deals with spatio-temporal models and data are summarized in the relevant task view (http://cran.r-project.org/web/views/SpatioTemporal.html) on the Comprehensive R Archive Network (CRAN) http://CRAN.R-project.org.

Second, in **3.2**, to show the predictive performance of TSCS. Specifically speaking, it refers to TSCS's high accuracy and good robustness when it making spatial interpolation.

Third, in **3.3**, to study some properties of TSCS referring to its predictive performance. This part of research aims at figuring out what factors affect the accuracy and robustness of TSCS interpolation, in order to offer us some rules of thumb on how to make a more effective prediction using TSCS according to different situations.

3.1 Data Generation

We decide to generate simulation data from 2D and 3D rectangular grid system separately. Each simulated data includes two parts – a complete historical spatio-temporal data (without missing value) and a new spatio-temporal data with missing observations to be interpolated. This data set is designed and generated resembling the observation set collected by sensors in the field of geostatistics.

3.1.1 2D Rectangular Grid System

First of all, we create a 2D rectangular grid system S regarded as the spatial domain of spatio-temporal data to be generated. Through combination of $x = 1, 2, ..., 50$ and $y = 1, 2, ..., 50$, we build a 2D rectangular grid system of 2500 spatial locations $S = \{s_1, s_2, ..., s_{2500}\}$, where $s_i = (x_i, y_i)$, $i \in \{1, 2, ..., 2500\}$ and $x_i, y_i \in \{1, 2, ..., 50\}$.

Next, we generate the overall spatio-temporal data through multiple time series of temporal domain $T = \{1, 2, ..., 650\}$ in all spatial locations $s_1, s_2, ..., s_{2500}$. The observed values of spatial locations are denoted by $z(s_1, t), z(s_2, t), ..., z(s_{2500}, t)$, $t \in T$. Before building time series model, we generate the initial value of time series for every spatial location with the following function.

$$C_i = z(x_i, y_i, t_0) : \begin{cases} r_i = f(x_i, y_i) = \frac{1}{100} x_i^2 + \frac{3}{100} y_i^2 - \frac{2}{5} x_i - \frac{1}{5} y_i + 5 \\ z_i = g(r_i) = \frac{\sin(r_i)}{r_i} \end{cases}$$
$$i = 1, 2, ..., 2500 \text{ and } x_i, y_i \in \{1, 2, ..., 50\}$$

Since we have obtained the initial value of time series for every spatial location, after careful consideration, we build the following time series model (Robert H. Shumway et al. 2015) to generate the overall spatio-temporal data.

$$z(s_i, t) = a_i \sin(\omega \cdot t) + b_i \cdot t + w_{it} + C_i, \quad i = 1, 2, ..., 2500 \text{ and } t = 1, 2, ..., 650$$

In this expression, C_i denotes initial value when $t = t_0$ and $a_i \sin(\omega \cdot t) + b_i \cdot t + w_{it}$ refers to function of time t in which $b_i \cdot t$ is linear trend term, $a_i \sin(\omega \cdot t)$ is periodic term and w_{it} is white noise satisfying $w_{it} \overset{iid}{\sim} N(0, \theta)$. Furthermore, values of these parameters are as follows. a_i is a random number from uniform distribution $U[1, 2]$, b_i is a random number from uniform distribution $U[0, 1/65]$, $\omega = 1/3$ and $\theta = 0.09$.

Here we split the overall spatio-temporal data into two parts. One is the historical spatio-temporal data $z(s_1, t), z(s_2, t), ..., z(s_{2500}, t)$, $t = 1, 2, ..., 500$. The other is the new spatio-temporal data $z(s_1, t), z(s_2, t), ..., z(s_{2500}, t)$, $t = 501, 502, ..., 650$. Within this new spatio-temporal data, we select its cross-section data of 7 different time points at regular intervals – t=501, t=525, t=550, t=575, t=600, t=625 and t=650. Meanwhile, 300, 500, 1000, 800, 400, 700 and 600 observations are randomly deleted respectively (2500 observations in total for each cross-section data). Finally, we write the historical spatio-temporal data and the coordinates of its spatial domain into CSV files **data1_2D.csv**. We also write the 7 cross-section data and the same coordinates into CSV files **newdata_2D.csv**.

3.1.2 3D Rectangular Grid System

In a similar way, we first create a 3D rectangular grid system S regarded as the spatial domain of spatio-temporal data to be generated. Through combination of $x = 1, 2, ..., 20$, $y = 1, 2, ..., 20$ and $h = 1, 2, ..., 20$, we build a 3D rectangular grid system of 8000 spatial locations $S = \{s_1, s_2, ..., s_{8000}\}$, where $s_i = (x_i, y_i, h_i)$, $i \in \{1, 2, ..., 8000\}$ and $x_i, y_i, h_i \in \{1, 2, ..., 20\}$.

Next, we generate the overall spatio-temporal data through multiple time series of temporal domain $T = \{1, 2, ..., 400\}$ in all spatial locations $s_1, s_2, ..., s_{8000}$. The observed values of spatial locations are denoted by $z(s_1, t), z(s_2, t), ..., z(s_{8000}, t)$, $t \in T$. Before building time series model, we generate the initial value of time series for every spatial location with the following function.

$$C_i = z(x_i, y_i, h_i, t_0) : \begin{cases} r_i = f(x_i, y_i, h_i) = x_i^2 + y_i^2 + h_i^2 \\ z_i = g(r_i) = \dfrac{\sin(r_i)}{r_i} \end{cases}$$

$$i = 1, 2, ..., 8000 \quad \text{and} \quad x_i, y_i, h_i \in \{1, 2, ..., 20\}$$

Since we have obtained the initial value of time series for every spatial location, after careful consideration, we build the following time series model (Robert H. Shumway et al. 2015) to generate the overall spatio-temporal data.

$$z(s_i, t) = a_i \sin(\omega \cdot t) + b_i \cdot t + w_{it} + C_i, \quad i = 1, 2, ..., 8000 \quad \text{and} \quad t = 1, 2, ..., 400$$

In this expression, C_i denotes initial value when $t = t_0$ and $a_i \sin(\omega \cdot t) + b_i \cdot t + w_{it}$ refers to function of time t in which $b_i \cdot t$ is linear trend term, $a_i \sin(\omega \cdot t)$ is periodic term and w_{it} is white noise satisfying $w_{it} \overset{\text{iid}}{\sim} N(0, \theta)$. Furthermore, values of these parameters are as follows. a_i is a random number from uniform distribution $U\left[\frac{3}{2}, 3\right]$, b_i is a random number from uniform distribution $U[0, 1/40]$, $\omega = 1/3$ and $\theta = 0.45^2$.

Here we split the overall spatio-temporal data into two parts. One is the historical spatio-temporal data $z(s_1, t), z(s_2, t), ..., z(s_{8000}, t)$, $t = 1, 2, ..., 300$. The other is the new spatio-temporal data $z(s_1, t), z(s_2, t), ..., z(s_{8000}, t)$, $t = 301, 302, ..., 400$. Within this new spatio-temporal data, we select its cross-section data of 6 different time points at regular intervals – t=301, t=320, t=340, t=360, t=380 and t=400. Meanwhile, 800, 1300, 2500, 2000, 1000 and 1800 observations are randomly deleted respectively (8000 observations in total for each cross-section data). Finally, we write the historical spatio-temporal data and the coordinates of its spatial domain into CSV files **data1_3D.csv**. We also write the 6 cross-section data and the same coordinates into CSV files **newdata_3D.csv**.

3.1.3 Explanations

For the avoidance of doubt, in this part, we make clear the reason why we choose the time series model in *3.1.1* and *3.1.2* together with parameter settings. Our considerations are as follows.

For one thing, characteristic of typical time series. A typical time series includes trend term, periodic trend (season) and noise. We choose additive model instead of multiplicative model because we hope the value of observation would not fluctuate drastically but change gradually. Besides, we use a simple function to generate data instead of ARIMA or GARCH model because these time series models primarily deal with stationary time series (usually difference of time series), but most of time series in the real world are nonstationary.

For another, the basic assumption – cointegrated system. Although a_i and b_i are set as random numbers within a given bound respectively, causing a small difference in magnitude of fluctuation, but the basic structure $(\sin(\omega \cdot t), t)$ of every time series is identical which leads to similar variation behavior. By doing that, through augmented Dickey–Fuller (ADF) test with significance level $\alpha = 0.01$, time series data of every spatial location is integrated of order 1. Furthermore, it also works giving the percentage of cointegrated relationships, a measurement of the degree our data satisfies the assumption of cointegrated system – 100%.

3.2 Workflow of TSCS

In this section, we demonstrate the workflow of spatial interpolation using **TSCS** package for 2D and 3D rectangular grid system respectively. Besides, the performance of TSCS handling simulated data is presented and evaluated, referring to its high accuracy and good robustness.

In the context of using **TSCS** package, the historical spatio-temporal data should be arranged in a standard format for input. As to 2D rectangular grid system, it should be a data frame containing these variables in order: X coordinate, Y coordinate and observations as time goes on. As to 3D rectangular grid system, it should be a data frame containing these variables in order: X coordinate, Y coordinate, Z coordinate and observations as time goes on. For this reason, data-preprocessing or data reconstruction is necessary beforehand.

In this package, the plotting functions for 2D and 3D cases are built upon **ggplot2** (Hadley Wickham and Winston Chang 2016) package and **rgl** (Daniel Adler and Duncan Murdoch 2017) package respectively.

3.2.1 2D Rectangular Grid System

Missing observations in new spatio-temporal data **newdata_2D** are shown in Fig 5.

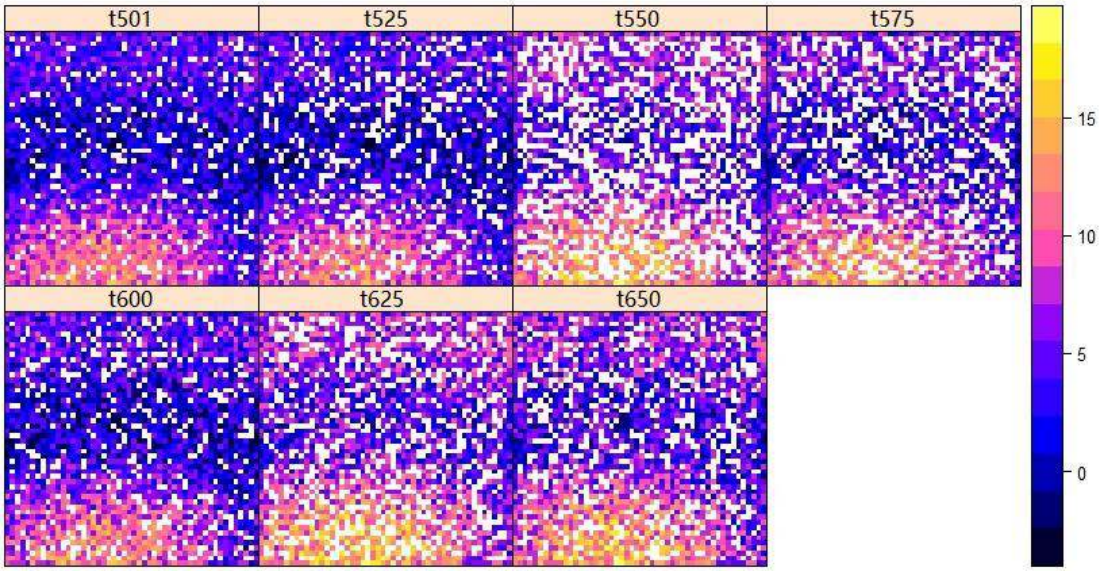

Fig 5. Missing observations in newdata_2D.

If you want to view the missing observations more clearly, **plot_NA** can help you do this.

Taking the cross-section data at time of t=501 for example, we visualize the spatial distribution of missing observations in Fig 6.

> plot_NA(newdata_2D[,c(1:2,3)])

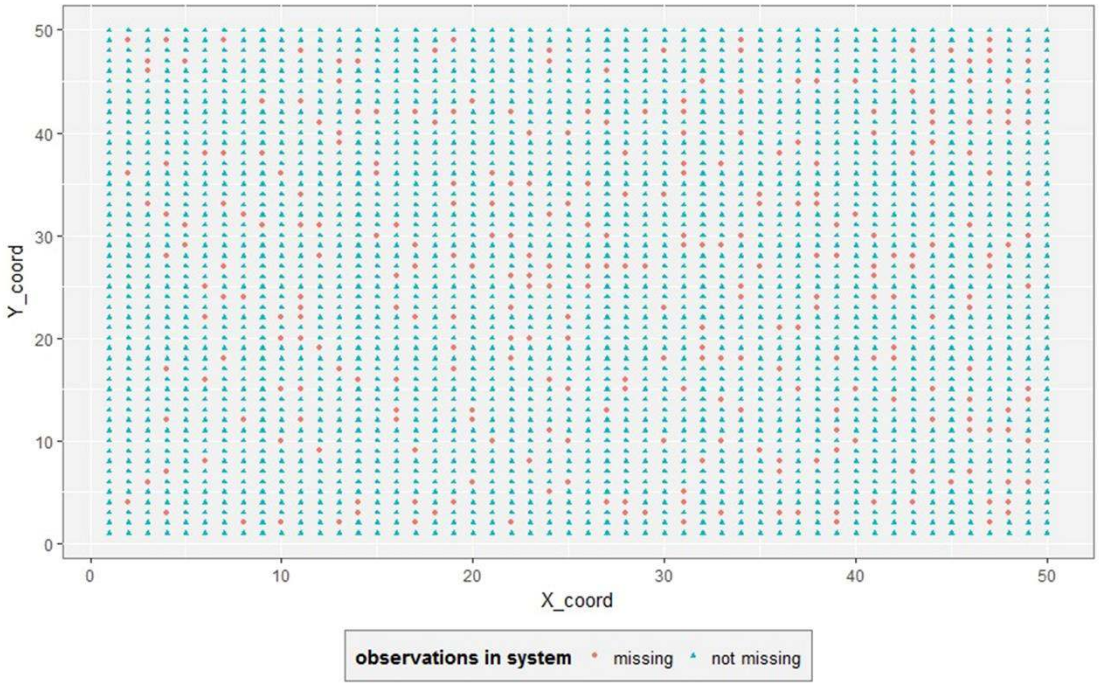

Fig 6. Missing observations in cross-section data at time of t=501.

plot_map draws two-dimensional spatial map with gradient color for a cross-section data. For instance, at time of t=501, the cross-section data are visualized in Fig 7.

> plot_map(newdata[,c(1:2,3)])

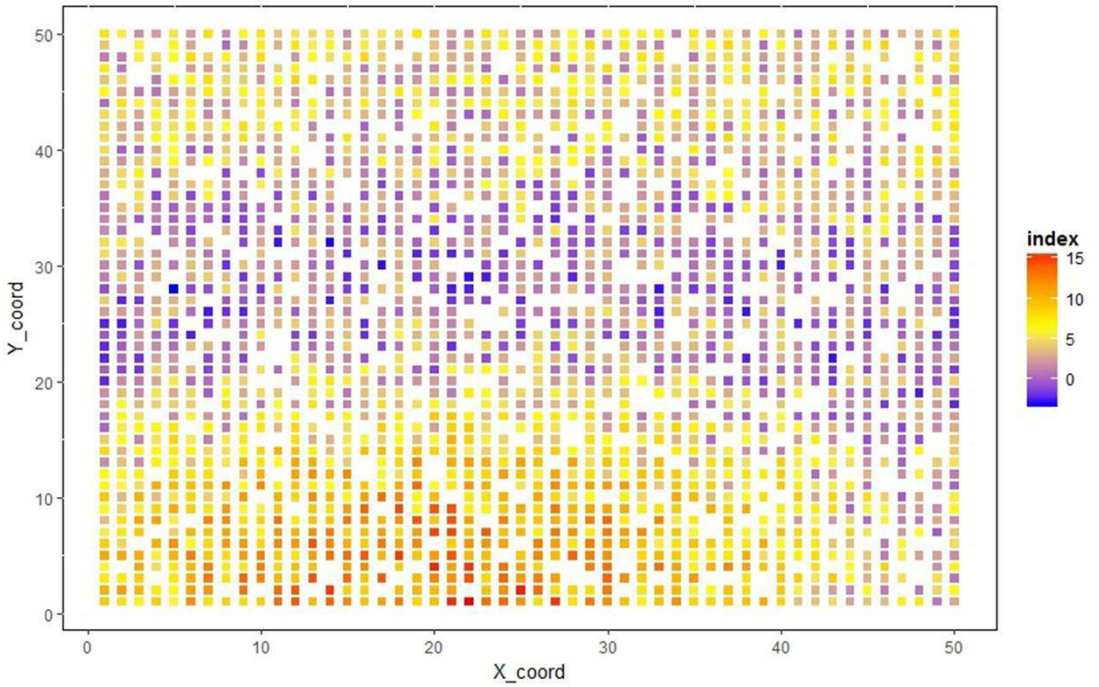

Fig 7. Spatial map of cross-section data at time of t=501.

Since we have an overview in mind of spatio-temporal data **newdata_2D**, we proceed to TSCS spatial interpolation. The first step, also the prerequisite, is obtaining regression

coefficient matrix with **tscsRegression** based on historical spatio-temporal data **data1_2D**. In function **tscsRegression**, the selection of adjacent spatial locations is carried out just as what Fig 8 shows.

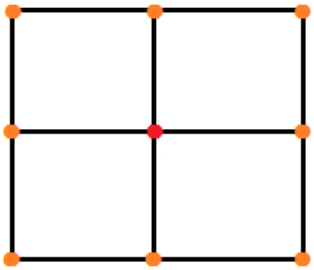

Fig 8. The way of selecting adjacent spatial locations for 2D rectangular grid system.
The red point is a given spatial location. The 8 yellow points are its adjacent spatial locations.

```
> basis <- tscsRegression(data = data1_2D, h = 1, v = 1, alpha = 0.01)
> basis$percentage
1
```

We can see that, with significance level 0.01, the percentage of cointegrated relationships is 100%, which means that the basic assumption of TSCS method is completely satisfied. This percentage is a measurement of the degree it satisfies the assumption of cointegrated system. It is highly affected by parameter **alpha**, the significance level you have set. Explicitly, smaller **alpha** results in smaller percentage.

Under cointegrated system **data1_2D** and **newdata_2D**, the work of estimating missing observations within **newdata_2D** can be done by utilizing **tscsEstimate**.

```
> est <- list()
> for (i in 3:9) {
+    est[[i-2]] <- tscsEstimate(matrix = basis$coef_matrix, newdata = newdata_2D[,c(1:2,i)],
                                h = 1, v = 1)
+ }
```

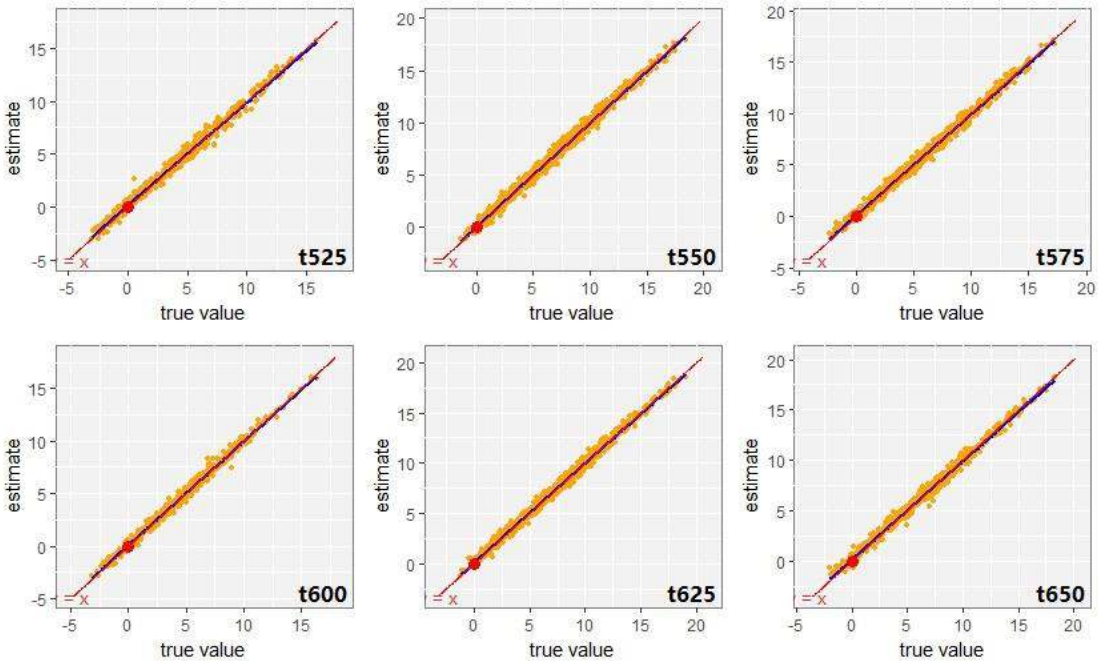

Fig 9. Graphic comparison between estimate and true value.

After spatial interpolation using TSCS method, provided that we have the true values of these missing observations saved in a list **trueValues**, a graphic comparison between true values and estimated values can be made by employing **plot_compare** (Fig 9). In this case, the result of TSCS interpolation is evaluated by two appraisal indexes – RMSE and standard deviation of error (they are clarified in the next part **3.3**), as summarized in Table 1.

Table 1. RMSE and standard deviation of error – evaluation of TSCS interpolation result.

–	t501	t525	t550	t575	t600	t625	t650
RMSE	0.3455	0.3924	0.3502	0.3604	0.3595	0.3589	0.3757
std	0.3452	0.3918	0.3498	0.3604	0.3595	0.3590	0.3752

3.2.2 3D Rectangular Grid System

As to 3D rectangular grid system, the procedure of spatial interpolation using **TSCS** is analogous to 2D case in *3.2.1*.

For new spatio-temporal data **newdata_3D**, **plot3D_NA** helps us view the missing observations more clearly. Taking the cross-section data at time of t=301 for example, we visualize the spatial distribution of missing observations in Fig 10 (A).

> plot3D_NA(newdata_3D[,c(1:3,4)])

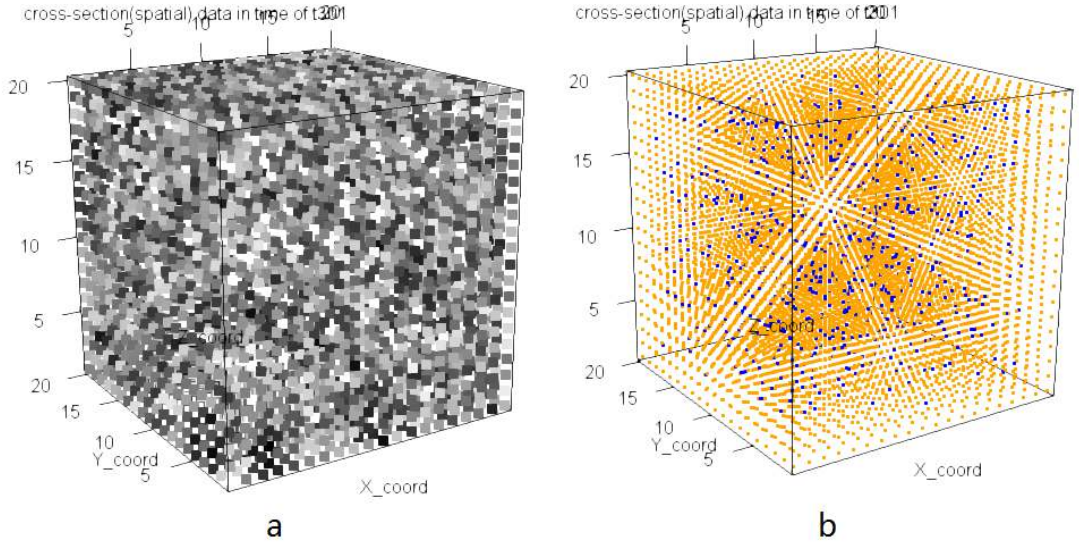

Fig 10. (a) Missing observations in cross-section data at time of t=301.
(b) Spatial map of cross-section data at time of t=301.

plot3D_map draws three-dimensional spatial map with gradient grey for a cross-section data. For instance, at time of t=301, the cross-section data are visualized in Fig 10 (B).

> plot3D_map(newdata[,c(1:3,4)])

With a basic overview of spatio-temporal data **newdata_3D**, we proceed to TSCS spatial interpolation. The first step is to obtain regression coefficient matrix through **tscsRegression3D** based on historical spatio-temporal data **data1_3D**. In function **tscsRegression3D**, the selection of adjacent spatial locations is carried out just as what Fig 11 shows.

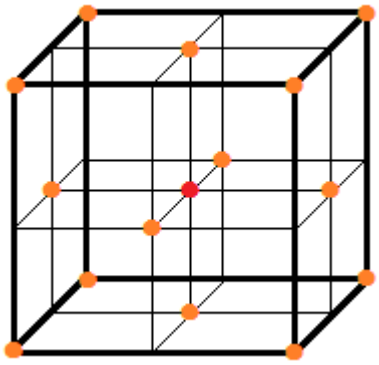

Fig 11. The way of selecting adjacent spatial locations for 3D rectangular grid system.
The red point is a given spatial location. The 14 yellow points are its adjacent spatial locations.

```
> basis <- tscsRegression3D(data = data1_3D, h1 = 1, h2 = 1, v = 1, alpha = 0.01)
> basis$percentage
1
```

Likewise, with significance level 0.01, the percentage of cointegrated relationships is 100%, which means that the basic assumption of TSCS method is completely satisfied. Under cointegrated system **data1_3D** and **newdata_3D**, estimation of missing observations within **newdata_2D** can be executed with the use of **tscsEstimate3D**.

```
> est <- list()
> for (i in 4:9) {
+    est[[i-3]] <- tscsEstimate3D(matrix = basis$coef_matrix, newdata = newdata_3D[,c(1:3,i)],
                                  h1 = 1, h2 = 1, v = 1)
+ }
```

After spatial interpolation with TSCS, if we have the true values of these missing observations saved in a list **trueValues**, the graphic comparison between true values and estimated values is arranged in Fig 12 (using **plot_compare**). And the evaluation of TSCS interpolation result is summarized in Table 2.

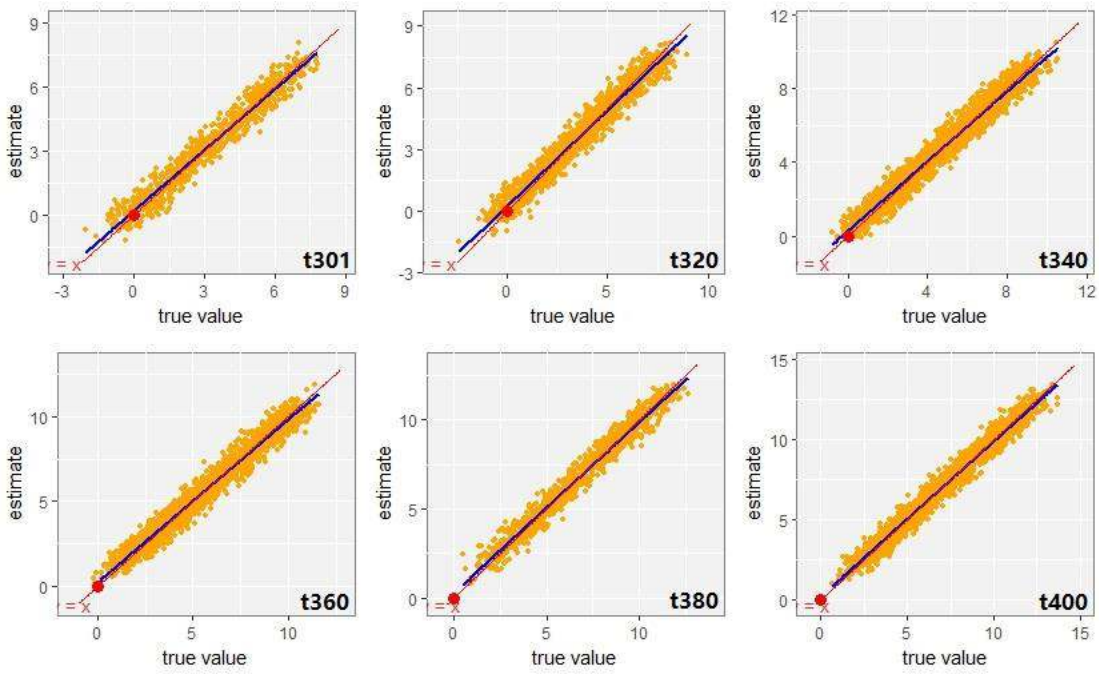

Fig 12. Graphic comparison between estimate and true value.

Table 2. RMSE and standard deviation of error – evaluation of TSCS interpolation result.

–	t301	t320	t340	t360	t380	t400
RMSE	0.4907	0.5009	0.4802	0.4994	0.4883	0.4816
std	0.4908	0.5007	0.4803	0.4986	0.4885	0.4809

3.3 Some Properties

The predictive ability of a model is of great concern. In the following, we study the three main influential factors for TSCS predictive performance. Our purpose of doing so is to provide us with insights into how to make more accurate estimations. Through repeated experiments of TSCS interpolation in a variety of cases, their results along with some important conclusions are summarized in *3.3.1 ~ 3.3.3*. For convenience but without loss of generality, the repeated experiments in this section are also based on the 2D spatio-temporal data **data1_2D, newdata_2D** observed at a 2D rectangular grid system and the 3D spatio-temporal data **data1_3D, newdata_3D** observed at a 3D rectangular grid system. They are generated in *3.1.1* and *3.1.2* respectively.

In this article, two appraisal indexes are recommended to evaluate the performance of TSCS interpolation quantitatively. The first is root-mean-square error (RMSE), used for measuring the differences between estimated values by a model and the values actually observed. Smaller RMSE means more accurate interpolation. The second is standard deviation of error, used for measuring how far the errors are spread out from their mean, namely, stability of errors. Smaller value means greater stability of errors, suggesting that errors would not fluctuate heavily due to difference of data.

3.3.1 Percentage of Missing Observation

From Table 3, we can conclude that, for 2D rectangular grid system, the percentage of missing observation in a cross-section data greatly influences the predictive performance of TSCS if the percentage is more than 60%. However, when the percentage is lower than 60%, TSCS's performance shows no much difference.

From Table 4, we find that there is almost no difference between values in the whole table, although both RMSE and standard deviation of error slightly increase with the percentage of missing observation increasing. It is obvious that, for 3D rectangular grid system, the robustness of TSCS method is much more better than its dealing with 2D rectangular grid system. The root cause is probably out of the selection of adjacent spatial locations. In three-dimensional cases, we select 14 adjacent spatial locations more than 8 in two-dimensional cases, thus including more information to explain the long-term equilibrium relationship that TSCS considers.

Table 3. RMSE and standard deviation of error for different percentage of missing observation, in regard to TSCS interpolation based on 2D spatio-temporal data **data1_2D** and **newdata_2D**.

Percentage	t501	t525	t550	t575	t600	t625	t650
4%	0.342	0.344	0.342	0.353	0.365	0.335	0.356
12%	0.351	0.355	0.339	0.349	0.372	0.332	0.369
20%	0.358	0.366	0.338	0.355	0.381	0.339	0.372
28%	0.362	0.361	0.342	0.361	0.377	0.341	0.371
36%	0.366	0.372	0.341	0.366	0.393	0.344	0.375
44%	0.380	0.389	0.346	0.375	0.425	0.356	0.390
52%	0.388	0.399	0.353	0.383	0.445	0.366	0.414

60%	0.454	0.464	0.377	0.449	0.582	0.398	0.509
68%	0.737	0.566	0.564	0.517	0.752	0.610	0.677
76%	0.563	0.638	0.444	0.550	0.747	0.495	0.622
84%	1.591	1.980	1.136	1.927	2.488	1.328	2.132
4%	0.339	0.345	0.342	0.353	0.364	0.336	0.355
12%	0.351	0.355	0.338	0.349	0.372	0.332	0.369
20%	0.358	0.365	0.338	0.355	0.381	0.339	0.372
28%	0.362	0.361	0.341	0.360	0.376	0.341	0.371
36%	0.366	0.372	0.340	0.365	0.393	0.343	0.375
44%	0.380	0.389	0.344	0.375	0.424	0.354	0.389
52%	0.388	0.398	0.349	0.383	0.444	0.364	0.413
60%	0.454	0.463	0.373	0.448	0.580	0.394	0.507
68%	0.736	0.564	0.556	0.516	0.751	0.601	0.675
76%	0.562	0.637	0.423	0.546	0.746	0.473	0.616
84%	1.588	1.978	1.071	1.910	2.486	1.256	2.100

Table 4. RMSE and standard deviation of error for different percentage of missing observation, in regard to TSCS interpolation based on 3D spatio-temporal data **data1_3D** and **newdata_3D**.

Percentage	t301	t320	t340	t360	t380	t400
10%	0.505	0.487	0.498	0.494	0.509	0.497
20%	0.504	0.494	0.499	0.498	0.509	0.498
30%	0.504	0.494	0.498	0.492	0.507	0.504
40%	0.506	0.500	0.499	0.492	0.507	0.502
50%	0.504	0.499	0.495	0.494	0.507	0.506
60%	0.508	0.501	0.499	0.496	0.508	0.507
70%	0.512	0.506	0.505	0.499	0.515	0.511
80%	0.522	0.519	0.516	0.512	0.525	0.526
10%	0.504	0.487	0.498	0.493	0.509	0.497
20%	0.504	0.494	0.499	0.497	0.509	0.497
30%	0.504	0.494	0.497	0.491	0.507	0.504
40%	0.506	0.500	0.499	0.491	0.505	0.501
50%	0.504	0.499	0.495	0.494	0.506	0.504
60%	0.508	0.501	0.499	0.495	0.507	0.505
70%	0.511	0.506	0.505	0.498	0.512	0.508
80%	0.522	0.519	0.516	0.510	0.521	0.520

3.3.2 Amount of Historical Spatio-Temporal Data

Based on results summarized in Table 5 and Table 6, it is easy to conclude that with the amount of historical spatio-temporal data increasing, TSCS interpolation is generally more accurate and more robust. It is in conformity with our practical experience that the more historical data you have used, the better the results of prediction are.

However, it is not always the case. The above conclusions are made based on a fundamental assumption that the system is relatively stable from past to future. As time goes, if some important properties of the system change greatly due to natural factors or human factors, the observed spatio-temporal data also changes a lot. Under this circumstance, a deep investigation needs to be carried out for figuring out how much historical data is usable.

Table 5. RMSE and standard deviation of error in regard to TSCS interpolation based on **newdata_2D** and different amount of historical spatio-temporal data from **data1_2D**.

Time Span	t501	t525	t550	t575	t600	t625	t650
t1~t500	0.374	0.380	0.343	0.372	0.399	0.351	0.388
t101~t500	0.379	0.387	0.346	0.381	0.415	0.360	0.404
t201~t500	0.386	0.396	0.356	0.400	0.440	0.385	0.440
t301~t500	0.389	0.407	0.381	0.433	0.486	0.450	0.514
t401~t500	0.378	0.418	0.433	0.502	0.587	0.597	0.685
t1~t500	0.374	0.380	0.342	0.372	0.399	0.351	0.387
t101~t500	0.379	0.387	0.345	0.381	0.415	0.359	0.403
t201~t500	0.385	0.395	0.354	0.400	0.440	0.384	0.439
t301~t500	0.389	0.407	0.380	0.432	0.486	0.449	0.514
t401~t500	0.378	0.418	0.431	0.502	0.586	0.596	0.685

Table 6. RMSE and standard deviation of error in regard to TSCS interpolation based on **newdata_3D** and different amount of historical spatio-temporal data from **data1_3D**.

Time Span	t301	t320	t340	t360	t380	t400
t1~t300	0.509	0.496	0.503	0.496	0.511	0.504
t51~t300	0.513	0.502	0.508	0.502	0.519	0.513
t101~t300	0.519	0.508	0.516	0.516	0.536	0.530
t151~t300	0.529	0.522	0.533	0.542	0.572	0.573
t201~t300	0.547	0.549	0.573	0.600	0.646	0.671
t1~t300	0.509	0.496	0.502	0.496	0.510	0.503
t51~t300	0.513	0.502	0.508	0.502	0.519	0.512
t101~t300	0.519	0.508	0.516	0.515	0.535	0.528
t151~t300	0.529	0.522	0.533	0.542	0.570	0.572
t201~t300	0.547	0.549	0.573	0.600	0.645	0.670

3.3.3 Step Length of Forward Prediction

As you see in Fig 13 and Fig 14, with the step length of forward prediction growing, the accuracy of TSCS interpolation gets worse, especially when the historical data is not abundant. It also corresponds with our rule of thumb that, for any prediction with uncertainty, the more steps ahead you choose, the less accurate prediction you make.

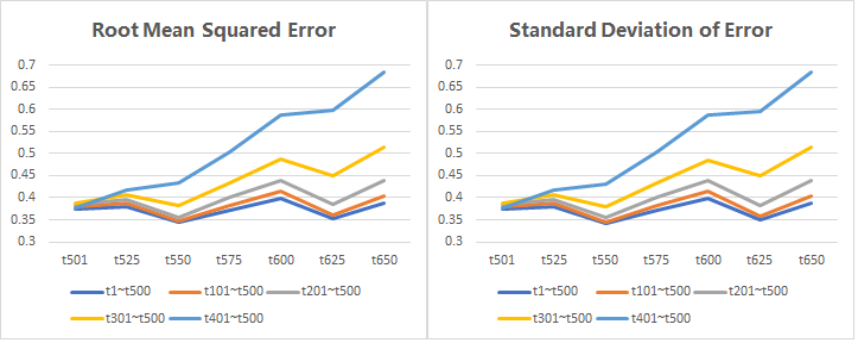

Fig 13. RMSE and standard deviation of error for different step length of forward prediction in regard to TSCS interpolation based on 2D spatio-temporal data data1_2D and newdata_2D.

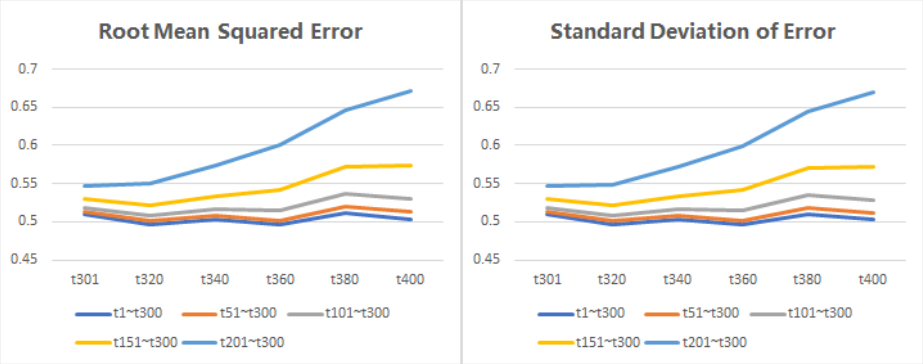

Fig 14. RMSE and standard deviation of error for different step length of forward prediction in regard to TSCS interpolation based on 3D spatio-temporal data data1_3D and newdata_3D.

4 Example: Analysis of GHCND Date Set

As what we have demonstrated in previous sections, TSCS method is simple and easy to use without model selection, parameter adjustment or requirement of subjective judgement. Moreover, it generally possesses high accuracy and good robustness when making interpolation. In most cases, if we have enough historical spatio-temporal data in hand, the main time-consuming work before TSCS interpolation is data pre-processing. Although it is a purely spatial interpolation method, these idiosyncrasies give it a chance to be regarded as a desirable alternative to existing spatio-temporal interpolation methods in some cases, where we merely intend to interpolate a series of cross-section data at each observed time point for a given spatial domain.

In this section, we are aimed at illustrating the strengths of TSCS in comparison with spatio-temporal kriging, one of the state-of-the-art spatio-temporal interpolation methods, in a real-world application based on Global Historical Climatology Network Data (GHCND) data sets. These strengths mainly refer to TSCS's good performance when dealing with a class of spatio-temporal interpolation problem introduced in the second paragraph of **Introduction** (please refer back to Fig 1).

4.1 Data Set

The real data that we select is named GHCND, which can be downloaded from URL ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/grid. The GHCND gridded dataset (HadGHCND) is produced through a joint effort between the United States National Oceanic and Atmospheric Administration (National Climatic Data Center) and the United Kingdom's Hadley Centre.

This big data set includes daily maximum temperature and minimum temperature from 1950 to 2016. The term *temperature* here does not mean real temperature but denotes temperature anomaly. The anomalies were calculated with respect to the following base period: 1961 to 1990. Hence, data of every year includes 2 data sets – *tmax* (daily maximum temperature anomaly) and *tmin* (daily minimum temperature anomaly). Besides, each of data set contains the following 6 columns (variables).

1st column: Month
2nd column: Day
3rd column: Grid box ID (value range: 1 to 7002, grid spacing = 3.75 deg×2.5 deg)
4th column: Longitude of lower left corner of grid box (degrees)
5th column: Latitude of lower left corner of grid box (degrees)
6th column: Temperature anomaly (whole degrees Celsius)

After appropriate data pre-processing for the purpose of implementing TSCS spatial interpolation using package **TSCS**, we decide to select data sets of daily maximum temperature anomaly from 2008 to 2012 as historical spatio-temporal data. Meanwhile, the new spatio-temporal data are 5 cross-section data selected in 2013 – 2013.1.1, 2013.4.2, 2013.7.2, 2013.10.1 and 2013.12.31.

The spatial domain of above data set is a 2D rectangular grid system, where every spatial location is a geo-spatial point in the world pinpointed by unique longitude and latitude. The distribution of these geo-spatial points covers almost the entire land area on the earth. (Fig 15,16,17)

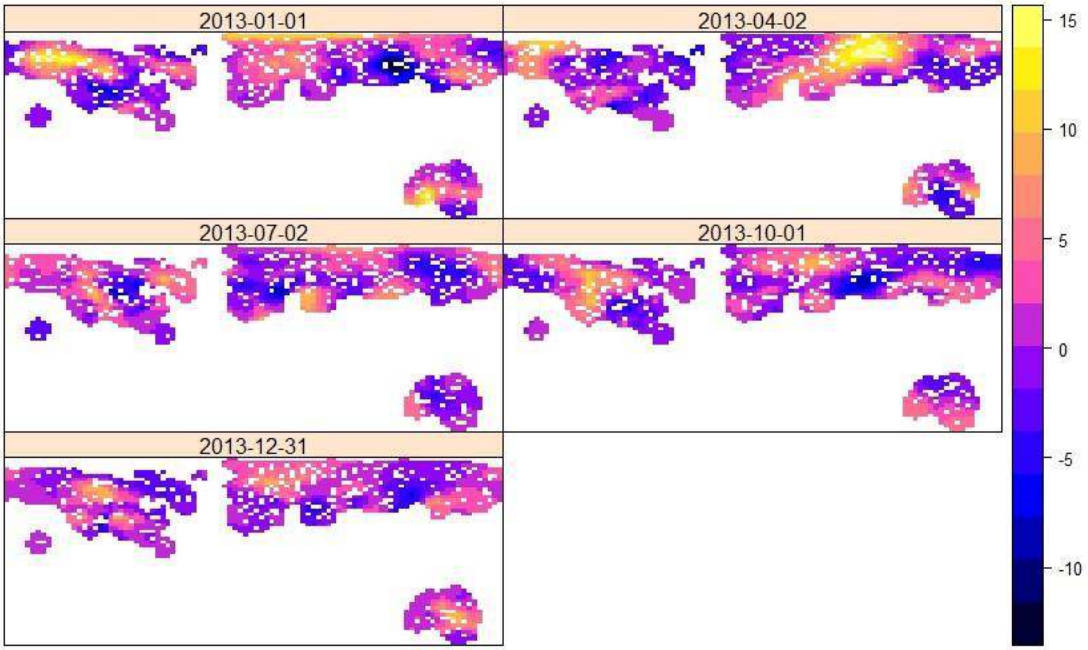

Fig 15. Missing observations in new spatio-temporal data.

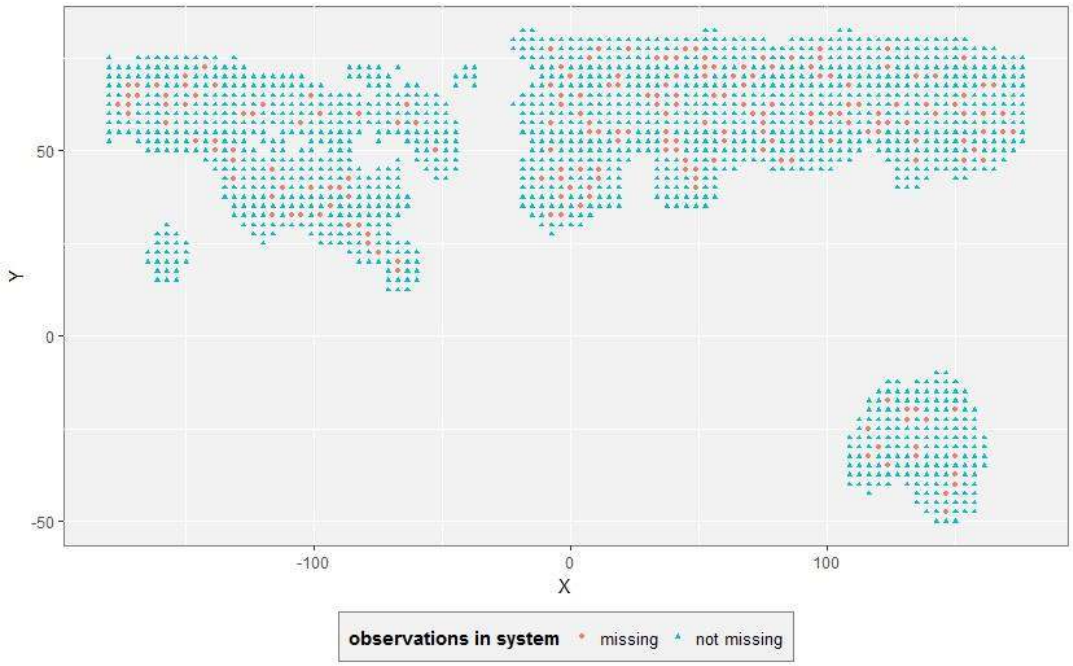

Fig 16. Missing observations in cross-section data on 2013.1.1 (using plot_NA).

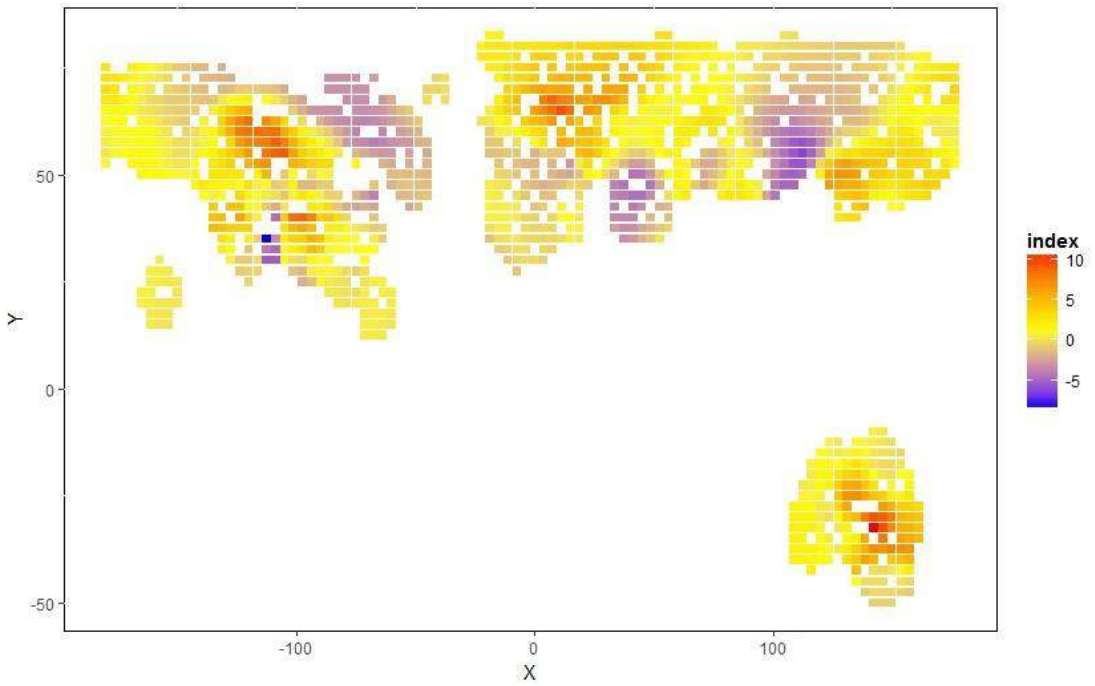

Fig 17. Spatial map of cross-section data on 2013.12.31 (using plot_map).

4.2 Spatio-Temporal Kriging

The following procedures about spatio-temporal kriging are all carried out by package **gstat**. Please refer to paper (Benedikt Graler et al. 2016) for more details.

Based on processed GHCND spatio-temporal data, the sample variogram is calculated and plotted in Fig18. After trying separable covariance model, product-sum covariance model, metric covariance model and sum-metric covariance model, the best fitting spatio-temporal variogram model is the product-sum covariance model, which can be identified from Table 7.

Table 7. Weighted MSE for different spatio-temporal variogram families and different choices for the one-dimensional variogram components. Columns denote the spatial and temporal variogram choices. The metric model is only applicable if both domains use the same family.

model	joint	Exp+Exp	Exp+Sph	Sph+Exp	Sph+Sph	Mat
separable	–	**10.22**	10.74	10.22	10.74	–
product-sum	–	4.13	5.30	**1.25**	2.67	–
metric	–	**3.31**	–	–	9.82	9.82
sum-metric	Exp	26.27	10.77	19.82	18.03	–
	Sph	**4.05**	4.22	5.10	4.45	–

A wireframe (3D) plot of sample variogram and the best fitting spatio-temporal variogram model in each family are presented in Fig 18.

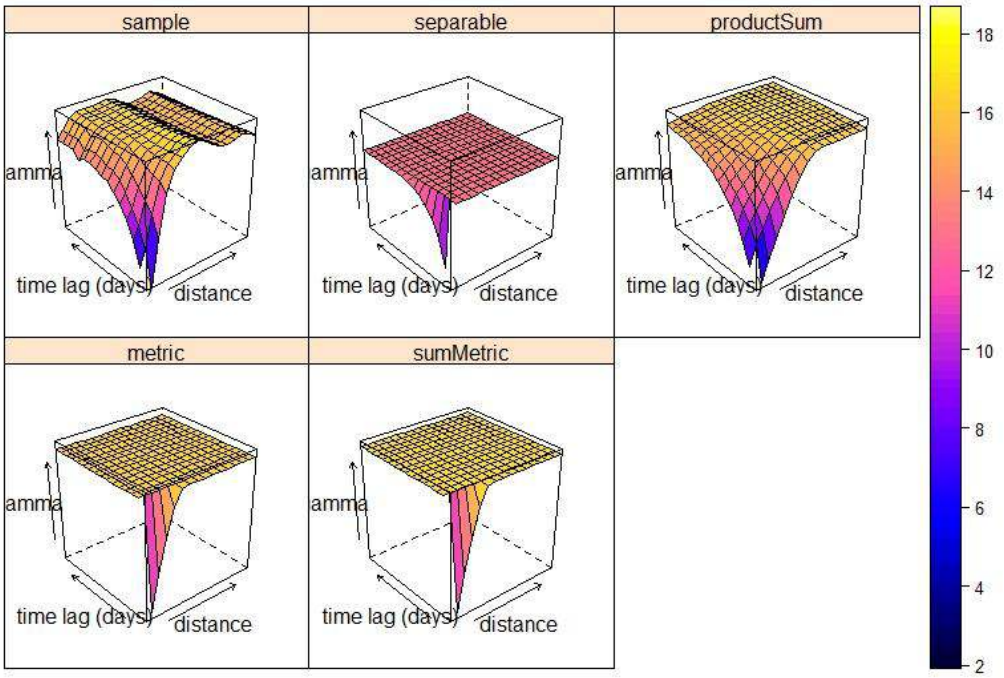

Fig 18. Sample variogram and fitted variogram models.

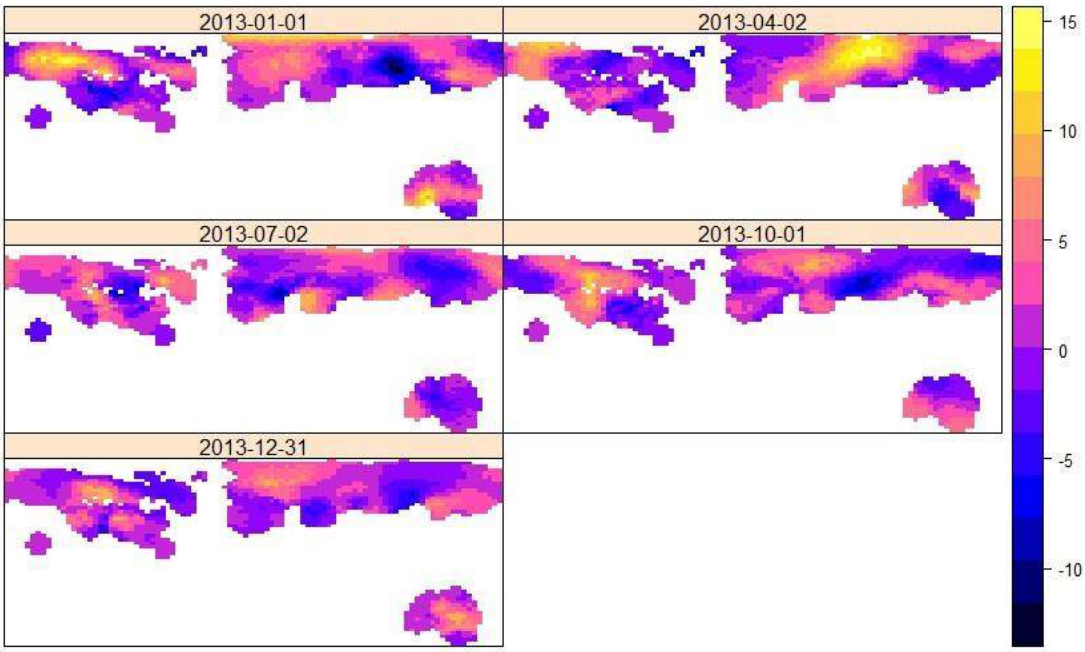

Fig 19. The full spatial map of data in 2013 after spatio-temporal interpolation using the product-sum covariance model.

As to spatio-temporal data of 2013, the full spatial map after interpolation using spatio-temporal kriging is presented in Fig 19. Moreover, since we have the true values of missing observations designated by white dots in Fig 15, a graphic comparison between true values and interpolation results are made for each spatial data on 2013.1.1, 2013.4.2, 2013.7.2, 2013.10.1 and 2013.12.31. They are shown in Fig 20 and the evaluation of TSCS interpolation result is summarized in Table 8.

Table 8 RMSE and standard deviation of error – evaluation of spatio-temporal kriging interpolation result.

–	2013.1.1	2013.4.2	2013.7.2	2013.10.1	2013.12.31
RMSE	1.6721	1.3078	1.2687	1.2353	1.1197
std	1.6750	1.3059	1.2567	1.2376	1.1216

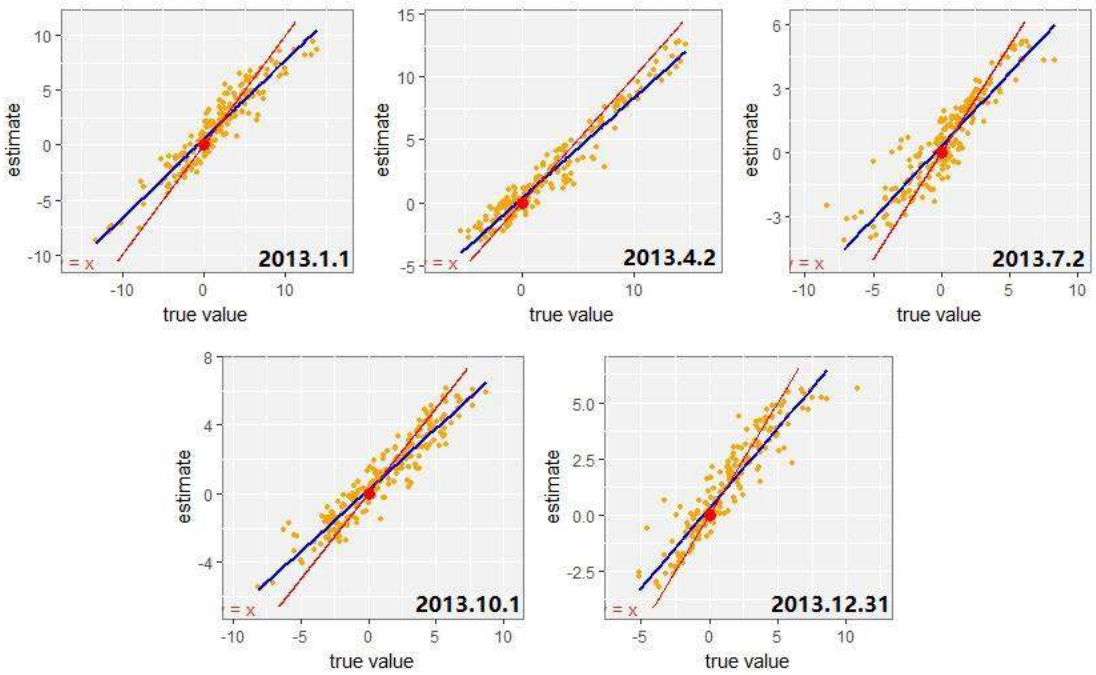

Fig 20. Graphic comparison between estimate and true value.

4.3 TSCS

Using **TSCS** package again, we interpolate the 5 cross-section data in the year of 2013. Likewise, the graphic comparison between true values and estimated values is shown in Fig 21 and the evaluation of TSCS interpolation result is summarized in Table 9.

Table 9 RMSE and standard deviation of error – evaluation of TSCS interpolation result.

–	2013.1.1	2013.4.2	2013.7.2	2013.10.1	2013.12.31
RMSE	0.3974	0.2783	0.3571	0.3289	0.4188
std	0.3983	0.2790	0.3560	0.3297	0.4183

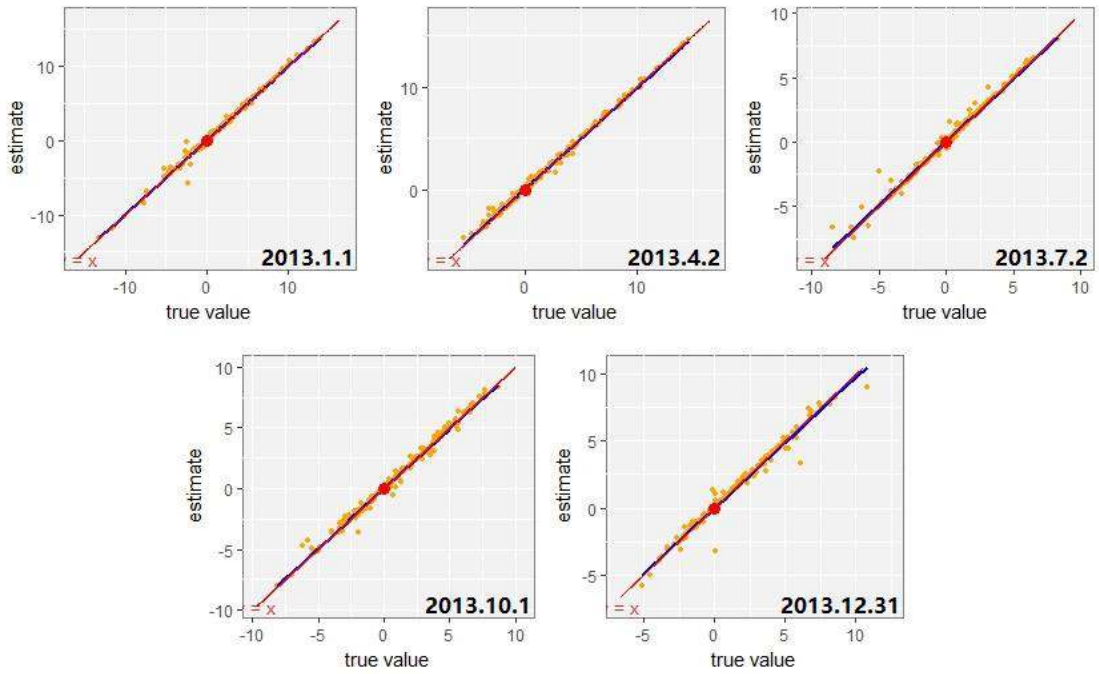

Fig 21. Graphic comparison between estimate and true value.

5 Discussion

5.1 Remarks

TSCS is not an ad hoc spatial interpolation method for certain specialized fields such as geostatistics but a general one. To some extent, TSCS is an original prototype to be modified or developed for more elaborate and specialized areas.

The selection of adjacent spatial locations is flexible indeed, not restricted to 8 neighbors for 2D rectangular grid system (Fig 8) or 14 neighbors for 3D rectangular grid system (Fig 11). In fact, any number of neighbors can be considered because the theory basis of TSCS is cointegrated relationship, but a more scientific way of selecting neighbors surely leads to more accurate interpolation. In the next major release of package **TSCS**, parameter **method** will be added to key functions – **tscsRegression, tscsRegression3D, tscsEstimate** and **tscsEstimate3D**, offering several alternatives of neighbor selection method. We will also add parameter **number** to them for the convenience of setting the number of neighbors freely. Theoretically, you can select more neighbors than normal or even all of the spatial locations in the whole map. This perhaps increase the accuracy of TSCS interpolation but it may be terribly time-consuming due to the limited computer performance currently. But then again, in most cases, 8 neighbors for 2D rectangular grid system (Fig 8) or 14 neighbors for 3D rectangular grid system (Fig 11) is enough to efficaciously cope with many spatial interpolation problems, but including more neighbors barely increases TSCS's accuracy, which is concluded from hundreds of simulated experiments.

TSCS is highly dependent on historical data, the historical spatio-temporal data. Hence, TSCS would be useless without it and TSCS will also be paralyzed if it is scarce. This is because the modelling of cointegrated relationship requires abundant time series data. For more details about what factors affect TSCS's performance, please refer back to **3.3**.

Last but not least, we need to make it explicit that the basic assumption of TSCS method – cointegrated system, is easily satisfied in the real world, even though it seems demanding. In a certain spatial domain, if the spatial locations distribute densely and the distance between

adjacent spatial locations is small enough, their variation trends in time series are often analogous so they would probably show strong correlations as time goes on, in other words, the long-term equilibrium relationship. In theory, long-term equilibrium relationship is equivalent to cointegrated relationship in some disciplines like econometrics. In further study, we want to research what adjustment we can make if the percentage of cointegration relationships fall short of 100% seriously and how this percentage, a property of data set, influences TSCS's performance.

5.2 Unsettled Problems

In regard to TSCS method, there is a minor defect both in theory and algorithm. Due to this defect, missing value is not allowed in historical spatio-temporal data. Therefore, we should take certain approaches to fill up these missing observations in advance, before making spatial interpolation with the package **TSCS**. It is a merely technical problem but requires further study with more effort on theory and its algorithmic details programmed with R.

TSCS of current version 0.1.1 is only able to handle spatio-temporal data collected on 2D or 3D rectangular grid system, two typical cases common in real life. Moreover, TSCS method of the current theory is only capable of interpolation but not extrapolation. It is unable to estimate missing observation located in the boundary or beyond the range of a given spatial domain. On balance, these problems result from theoretical flaws which remain to be further studied. These unsettled problems are expected to be solved in the next major release of the package **TSCS**.

References

Alok Bhargava (1986) On the theory of testing for unit roots in observed time series. Rev Econ Stud 53(3):369-384

Anindya Banerjee, Juan J. Dolado, John W. Galbraith, David Hendry (1993) Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data. Oxford University Press, pp 70–81

Benedikt Graler, Edzer Pebesma, Gerard Heuvelink (2016) Spatio-Temporal Interpolation using gstat. CRAN. https://cran.r-project.org/web/packages/gstat/vignettes/spatio-temporal-kriging.pdf

Daniel Adler, Duncan Murdoch (2017) 3D Visualization Using OpenGL. https://r-forge.r-project.org/projects/rgl/. R package version 0.98.1

Donald E. Myers. What Is Geostatistics? http://www.u.arizona.edu/~donaldm/homepage/whatis.html.

Felix Cheysson (2016) Modelling Space Time AutoRegressive Moving Average (STARMA) Processes. https://CRAN.R-project.org/package=starma. R package version 1.3.

Hadley Wickham, Winston Chang (2016) Create Elegant Data Visualisations Using the Grammar of Graphics. http://ggplot2.tidyverse.org/. R package version 2.2.1

Hua Xu, Qiang Wu, Hongzhuan Lei, Shiyong Li (2012) Study on spatial interpolation method and its application. 2012 International Conference on Audio, Language and Image Processing, pp 1057-1061

Johan Lindstrom, Adam Szpiro, Paul D. Sampson, Silas Bergen, Lianne Sheppard (2013) SpatioTemporal: An R Package for Spatio-Temporal Modelling of Air-Pollution. CRAN. https://cran.r-project.org/web/packages/SpatioTemporal/vignettes/ST_intro.pdf

Martin Schlather, Alexander Malinowski, Peter J. Menck, Marco Oesting, Kirstin Strokorb (2015) Analysis, Simulation and Prediction of Multivariate Random Fields with Package RandomFields. J STAT SOFTW 63(8):1-25

Michael H. Kutner, Christopher J. Nachtsheim, John Neter (1988) Applied Linear Regression

Models. McGraw-Hill/Irwin, New York
Robert H. Shumway, David S. Stoffer (2015) Time Series Analysis and Its Applications. Springer New York Dordrecht Heidelberg London
Shiying Zhang, Zhi Fan, Mingyuan Guo (2014) Cointegration Theory and Volatility Model. Tsinghua University Press, Beijing
Stein, M. L. (1999). Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York