Journal of the Korean Statistical Society 46 (2017) 641–653

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Multiscale representation for irregularly spaced data

Dongik Jang [a], Donghoh Kim [b], Kyungmee O. Kim [c,*]

[a] Department of Transport Big Data Research, The Korea Transport Institute, Sejong 30147, Republic of Korea
[b] Department of Applied Mathematics, Sejong University, Seoul 05006, Republic of Korea
[c] Department of Industrial Engineering, Konkuk University, Seoul 05029, Republic of Korea

ARTICLE INFO

Article history:
Received 24 July 2017
Accepted 7 September 2017
Available online 9 October 2017

AMS 2000 subject classifications:
62G08
62H12

Keywords:
Irregularly spaced data
Multiscale method
Pseudo data
Smoothing splines
Thin-plate splines
Wavelets

ABSTRACT

In this paper, we propose a multiscale method for representing inhomogeneous functions (surfaces) from irregularly spaced noisy observations that have inherent multiscale structure. The proposed multiscale method is based on a novel combination of the standard discrete wavelet transform with newly defined pseudo data. The pseudo data, which can be considered as a preprocessing of the original data, play a crucial role in deriving the proposed method and motivating a practical algorithm. The proposed algorithm using the empirical pseudo data is computationally fast, simple to describe and easy to implement. Moreover, results from numerical examples and real data analysis demonstrate the promising empirical properties of the proposed method.

1. Introduction

1.1. Background and motivation

Suppose that we are interested in analysis of apartment prices in Seoul metropolitan area shown in Fig. 1. More specifically, we want to construct a movie of the apartment price variability and then to represent spatial–temporal patterns of the prices through the movie. A better understanding of spatial–temporal patterns of the prices is crucial for evaluating government policy of housing market and building a housing pricing model.

To accomplish this goal, it is necessary to estimate the underlying field of apartment prices effectively. As shown in Fig. 1, the price data are scattered, and they have inhomogeneous spatial densities including data voids of various sizes, that is, the global trend of the data is coupled with local anomalies of different size of areas. In order to represent various spatial patterns of the prices, it is required to have a multiscale representation that examines the important features of the data according to various spatial scales. Hence, we emphasize that main concern of this study is beyond a simple regression contest. An ideal method for a successful study holds the following properties:

(a) It will, given scattered data, be efficient to estimate the underlying field of inhomogeneous spatial data where the data density varies considerably over the area.
(b) It will hold a multiscale structure that is capable of expressing spatial dynamics of the data.
(c) It will be computationally fast so that it can be applied in a 'data mining' context to massive spatial–temporal data.

* Corresponding author.
E-mail address: kyungmee@konkuk.ac.kr (K.O. Kim).

https://doi.org/10.1016/j.jkss.2017.09.002
1226-3192/

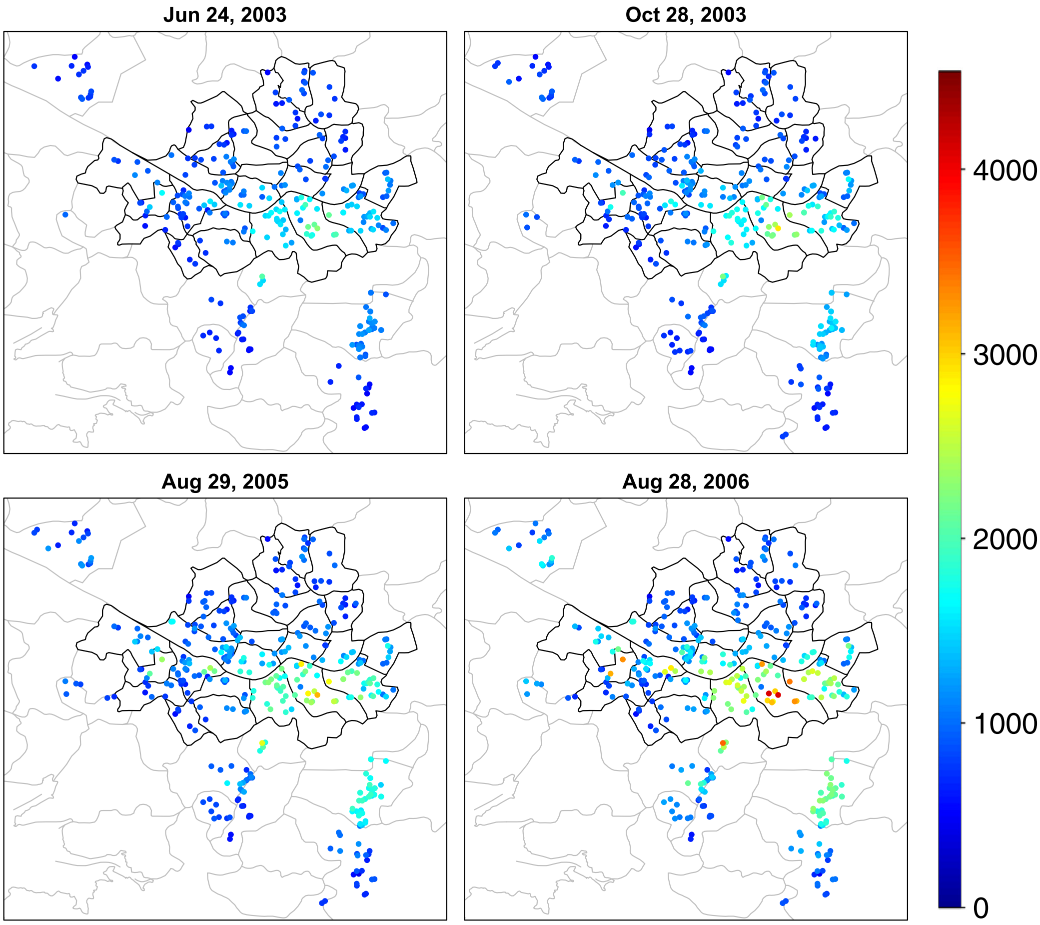

Fig. 1. The prices and locations of apartment complex in Seoul and its suburban area. The numbers in the legend denote Korean won($\times 10^4$) per 3.3 m^2.

Most popular multiscale approach for satisfying the above properties may be wavelet shrinkages that have excellent theoretical and empirical properties for a large class of the functions. However, two major assumptions are made when implementing standard discrete wavelet transform-based methods for a dataset: (i) the number of observations, say n, is dyadic, i.e., $n = 2^J$ for some integer J, and (ii) the data are observed on a regular and squared grid. Although conventional wavelet shrinkage methods enjoy a fast and efficient algorithm, the above mentioned assumptions strongly reduce the applicability of wavelet shrinkages to various data such as the data in Fig. 1.

We remark that the apartment price data we used are provided by Kookmin Bank, which are obtained by sampling through their network. The dataset consists of weekly apartment prices by apartment complex and size of the apartment. Originally, the data cover the apartment prices of the whole country. In this study, we consider only Seoul and its suburban area, for the apartment prices of these regions are of primary interest to the economic policy makers and common people.

1.2. Existing methods and our contribution

Given n pairs of observations (x_i, z_i), $i = 1, \ldots, n$, we assume an additive model satisfying

$$z_i = f(x_i) + \epsilon_i, \tag{1}$$

where ϵ_i's are independent and identically distributed random errors with mean zero and finite variance σ^2 and f is an unknown function of interest that might be spatially inhomogeneous. For notational simplicity, $f_i := f(x_i)$ is the ith element of $\boldsymbol{f} = \boldsymbol{W}^T\boldsymbol{\theta}$, $\boldsymbol{W}$ is orthogonal wavelet operator, and $\boldsymbol{\theta}$ is the vector of the wavelet coefficients of $\boldsymbol{f}$. Various attempts have been made to relax the requirements of a sample size being a power of 2 and the designs being regularly spaced. Two types of approaches are used to overcome the above problems. The first approach is to transform the original data z into new data z^* which satisfies the assumptions, so that z^* can undergo the standard wavelet shrinkage. Along with this approach, many methods have been developed based on transformation (Cai & Brown, 1998; Pensky & Vidakovic, 2001),