Spatial and Spatio-temporal Epidemiology 000 (2016) 1–11

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Accelerating the discovery of space-time patterns of infectious diseases using parallel computing

Alexander Hohl[a], Eric Delmelle[a,*], Wenwu Tang[a], Irene Casas[b]

[a] Department of Geography and Earth Sciences and Center for Applied GIScience, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[b] School of History and Social Sciences, Louisiana Tech University, Ruston, Louisiana, 71272, USA

ARTICLE INFO

Article history:
Received 16 November 2015
Revised 3 May 2016
Accepted 17 May 2016
Available online xxx

Keywords:
Dengue fever
Parallel computing
Space-time analysis

ABSTRACT

Infectious diseases have complex transmission cycles, and effective public health responses require the ability to monitor outbreaks in a timely manner. Space-time statistics facilitate the discovery of disease dynamics including rate of spread and seasonal cyclic patterns, but are computationally demanding, especially for datasets of increasing size, diversity and availability. High-performance computing reduces the effort required to identify these patterns, however heterogeneity in the data must be accounted for. We develop an adaptive space-time domain decomposition approach for parallel computation of the space-time kernel density. We apply our methodology to individual reported dengue cases from 2010 to 2011 in the city of Cali, Colombia. The parallel implementation reaches significant speedup compared to sequential counterparts. Density values are visualized in an interactive 3D environment, which facilitates the identification and communication of uneven space-time distribution of disease events. Our framework has the potential to enhance the timely monitoring of infectious diseases.

1. Introduction

1.1. Detecting disease outbreaks

Infectious diseases have complex transmission cycles, and an effective public health response requires the ability to monitor and analyze outbreaks under critical space-time conditions (Eisen and Eisen, 2011). Space-time analytics and geovisualization are particularly attractive to analyze spatial data with a time stamp (Jacquezet al., 2005; Rogerson and Yamada, 2008; Robertson et al., 2010; Kulldorff, 2010), as they facilitate the discovery of inherent patterns (rate of disease spread, cyclic pattern, direction, intensity and risk of diffusion to new regions). The identification of a cluster of illness provides critical intelligence for response; timely and focused monitoring is thus a critical element to reduce the burdens associated with diseases. The detection of space-time clusters can be computationally demanding, and this issue is exacerbated with spatiotemporal datasets of increasing size, diversity and availability (Grubesic et al., 2014; Robertson et al., 2010). Accelerated processing capabilities are therefore critical to reduce the computational effort when conducting space-time analysis on epidemiological datasets, and particularly so at the individual level. However, careful spatiotemporal domain decomposition is often necessary to prevent workload imbalances, thereby reducing computing inefficiency. Heterogeneity in the data has to be accounted for when designing new algorithms capable of implementing parallel strategies and integrating time along spatial dimension.

A series of statistical approaches has been used to detect spatial or spatiotemporal clusters of infectious diseases. The Knox test for space-time interaction evaluates

* Corresponding author. Tel.: +1 7046875991.
E-mail address: eric.delmelle@uncc.edu (E. Delmelle).

http://dx.doi.org/10.1016/j.sste.2016.05.002
1877-5845/

the presence of a space-time cluster at given spatial and temporal distances (Kulldorff and Hjalmars, 1999). Knox' method is limited due to its arbitrary definition of closeness (Robertson et al., 2010), and the Mantel's test (Mantel, 1967) incorporates the notion of distance decay in that nearby pairs of events are more important than distant pairs. The space-time Ripley's K function evaluates the magnitude of space-time clustering at different spatial and temporal scales (Bailey and Gatrell, 1995). The spatial scan statistic (Kulldorff, 1997) identifies the most likely disease clusters in a study area by maximizing the likelihood that disease cases are located within a set of concentric circles that are moved across the study area. In a space-time context, the scan statistic uses a cylinder instead of a circle, where the vertical axis represents time (Kulldorff et al. 2005).

In addition to scan statistics, autocorrelation-based methods have been widely used for identifying clusters in epidemiology (Mclafferty, 2015). Global methods, such as Moran's I (Moran, 1950), tell us whether clustering of similar attribute values is present within the study area, while its local version (local Moran's *I*) identifies locus and shape of these clusters (Anselin, 1995). Finally, kernel density estimation (KDE) techniques are used to generate continuous surfaces of disease intensity (Carlos et al. 2010; Delmelle et al. 2014b). The temporal extension of the KDE is known as the space-time kernel density estimation (STKDE) and essentially maps a volume of disease intensity along the space-time domain (Nakaya and Yano, 2010). However, the above methods are computationally intensive (Robertson et al. 2010), especially when Monte Carlo simulations are used for significance testing, and when the temporal dimension is added, resulting in long execution times when compared to their planar counterparts (Costa et al., 2012). Expedited processing capabilities are critical for analyzing spatiotemporal epidemiological data of increasing size, diversity and availability. High-performance and parallel computing offer the capacity to solve computationally demanding problems in limited time frame.

1.2. Parallel computing

Parallel computing is based on a divide and conquer strategy that breaks down a problem into sub-problems that are small enough to be handled computationally. Then, the solutions to the sub-problems are aggregated to form a solution for the original problem (Wilkinson and Allen, 2004). The general approach for parallel computing is to decompose a dataset into smaller subsets, for example, along the spatial or temporal dimension, distribute the resulting subdomains to multiple processors for parallel processing, and finally collect and reassemble the results (Ding and Densham, 1996). However, to prevent workload imbalance among processors and, therefore, processing inefficiency, the spatially explicit characteristics of the data often need to be accounted for (Wang and Armstrong, 2003). While random or uniform data can be decomposed by non-adaptive regular tessellations, doing so for clustered datasets results in heterogeneous subdomains as they contain uneven quantities of data (Ding and Densham, 1996). Such subdomains exhibit variation in computational intensities (Wang and Armstrong, 2009) resulting in workload imbalance and inefficiency. Recursive domain decomposition methods, such as quadtrees, have been widely used for mitigating workload imbalance, especially for spatially heterogeneous data (Turton, 2003; Wang and Armstrong, 2003).

Despite the recent popularity of incorporating time dimension in geographic models (Kwan and Neutens, 2014), the recursive decomposition of spatiotemporal datasets has been insufficiently addressed in the literature. Most of the work reported is based on static spatial domain decomposition. For example, spatial domain decomposition has been used for parallel computation of the $G_i^*(d)$ statistic (Armstrong and Marciano, 1995; Wang et al., 2008). Liu and colleagues compared non-adaptive with adaptive domain decomposition for parallel processing, and found that adaptive decomposition often leads to increase in workload balance (Liu et al. 2010). Spatial domain decomposition, including both regular (block) and ordered (cyclic; taking into account heterogeneity in spatial data) strategies, has been applied to parallelize the AMOEBA algorithm for detection of spatial clustering patterns (Widener et al., 2012). Static spatial domain decomposition has been used to compute space-time kernel density (STKDE) in parallel on reported dengue fever cases (Delmelle et al. 2014a). The latest efforts of accelerating clustering algorithms include the use of general-purpose graphics processing units (GPGPU), for instance, to select optimal bandwidths for kernel density estimation (Andrzejewski et al., 2013), and to compute the Ripley's K function on massive spatial point data (Tang et al., 2015).

1.3. Objectives

In this article, we develop a parallel computing approach based on adaptive space-time domain decomposition for the acceleration of the space-time kernel density estimation (STKDE), a computationally demanding space-time statistic. We apply our methodology on an epidemiological dataset of reported dengue fever cases in the city of Cali, Colombia for the years 2010 and 2011. We implement a parallel computing approach to conduct the space-time K function test, on both observed cases and population adjusted Monte Carlo simulations. Optimal space-time K-function parameters serve as inputs for the STKDE. Based on the K-function analysis, we choose the spatial and temporal bandwidths for STKDE.

2. Materials and methods

2.1. Data and study area

The city of Santiago de Cali (from here on referred to as Cali) is located in the valley of the Cauca River; limited to the east by the river and to the west by a mountain system. It has a tropical climate with two distinct rainy seasons, from April to July and September to December. The yearly average temperature is 26 °C (79 °F) which results in perfect conditions for the Aedes Aegypti mosquito to reproduce (de Cali, 2008). The city of Cali with a population of around 2.5 million people is considered an endemic

Please cite this article as: A. Hohl et al., Accelerating the discovery of space-time patterns of infectious diseases using parallel computing, Spatial and Spatio-temporal Epidemiology (2016), http://dx.doi.org/10.1016/j.sste.2016.05.002