Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 25 (2009) 498–517
international journal of forecasting
www.elsevier.com/locate/ijforecast

Expectation-based scan statistics for monitoring spatial time series data

Daniel B. Neill*

H.J. Heinz III College, School of Public Policy and Management, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States

Abstract

We consider the simultaneous monitoring of a large number of spatially localized time series in order to detect emerging spatial patterns. For example, in disease surveillance, we detect emerging outbreaks by monitoring electronically available public health data, e.g. aggregate daily counts of Emergency Department visits. We propose a two-step approach based on the *expectation-based scan statistic*: we first compute the expected count for each recent day for each spatial location, then find spatial regions (groups of nearby locations) where the recent counts are significantly higher than expected. By aggregating information across multiple time series rather than monitoring each series separately, we can improve the timeliness, accuracy, and spatial resolution of detection. We evaluate several variants of the expectation-based scan statistic on the disease surveillance task (using synthetic outbreaks injected into real-world hospital Emergency Department data), and draw conclusions about which models and methods are most appropriate for which surveillance tasks.

Keywords: Time series monitoring; Pattern detection; Event detection; Spatial scan statistics; Biosurveillance

1. Introduction

Many applications require the monitoring of time series data in order to detect anomalous counts. A traditional application of time series monitoring is the use of statistical process control to ensure consistency in manufacturing: the process is measured regularly to ensure that the desired specifications (e.g. product size and weight) remain within an acceptable range. More recently, time series monitoring has been used in a variety of event detection systems: crime surveillance systems (Gorr & Harries, 2003; Levine, 1999) detect emerging hot-spots of crime activity, disease surveillance systems (Sabhnani et al., 2005) monitor electronic public health data such as hospital visits and medication sales in order to detect emerging outbreaks, and environmental monitoring systems (Ailamaki, Faloutsos, Fischbeck, Small, & VanBriesen, 2003) detect abnormally high pollutant levels in the air, water, and soil.

* Tel.: +1 412 268 3885.
E-mail address: neill@cs.cmu.edu.

0169-2070/$ - see front matter © 2008 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2008.12.002

In all of these event detection applications, we wish to detect emerging spatial patterns as quickly and accurately as possible, enabling a timely and appropriate response to the detected events. As a concrete example, we focus on the task of detecting outbreaks of respiratory illness using hospital Emergency Department (ED) data. In this case, we can monitor the number of patients visiting the ED with respiratory symptoms in each zip code on each day. Each zip code s_i has a corresponding time series of daily counts c_i^t, and our goal is to detect anomalous increases in counts that correspond to an emerging outbreak of disease.

A variety of methods have been developed to monitor time series data and detect emerging anomalies. Control chart methods (Shewhart, 1931) compare each observed count to its expected value (a counterfactual forecast obtained from time series analysis of the historical data), and detect any observations outside a critical range. Cumulative sum methods (Page, 1954) and tracking signals (Brown, 1959; Trigg, 1964) aggregate these deviations across multiple time steps in order to detect shifts in a process mean. When extending these techniques to the simultaneous monitoring of multiple time series, we have several options (Burkom, Murphy, Coberly, & Hurt-Mullen, 2005). In the simplest, "parallel monitoring" approach, we monitor each time series separately and report any anomalous values. In the "consensus monitoring" approach, we combine the signals from multiple time series in order to achieve higher detection power. To detect anomalies that affect multiple time series simultaneously, we can either combine the outputs of multiple univariate detectors or treat the multiple time series as a single multivariate quantity to be monitored. For example, multivariate control charts (Hotelling, 1947) learn the joint distribution of a set of signals from historical data, and detect when the current multivariate signal is sufficiently far from its expectation.

We note, however, that none of these time series monitoring methods account for the *spatial* nature of the event detection problem. We expect events to be localized in space: if a given location is affected by the event, nearby locations are more likely to be affected than locations that are spatially distant. For example, disease outbreaks tend to affect spatially contiguous areas, either because of contagion (e.g. human-to-human transmission) or because the cases share a common source (e.g. contaminated drinking water). Thus, we must consider alternate methods of monitoring spatial time series data, where we expect anomalies to affect the time series for some spatially localized subset of locations.

A typical approach to the monitoring of spatial time series data uses "fixed partitions": we map the locations to a Euclidean space (e.g. using the longitude and latitude of each zip code centroid), partition the search space such that each location is contained in exactly one partition, and aggregate the counts for each partition into a single time series. We then monitor the time series for each partition separately, and report any anomalous counts. One challenge is deciding how to partition the search space: in the case of zip code level data, we could consider each zip code to be a separate partition, combine multiple adjacent zip codes in a single partition, or even aggregate all of the zip codes into a single time series. An alternative is to form an "ad-hoc partitioning" by identifying individual locations with high counts and using some heuristic to cluster these locations (Corcoran, Wilson, & Ware, 2003).

Any choice of partitioning scheme creates a set of potential problems, which we call the "curse of fixed partitions". In general, we do not have *a priori* knowledge of how many locations will be affected by an event, and we wish to maintain high detection power whether the event affects a single location, all locations, or anything in between. A coarse partitioning of the search space will lose power to detect events that affect a small number of locations, since the anomalous time series will be aggregated with other counts that are not anomalous. A fine partitioning of the search space will lose power to detect events that affect many locations, since only a small number of anomalous time series are considered in each partition. Partitions of intermediate size will lose some power to detect both very small and very large events. Moreover, even if the partition size corresponds well to the event size, the fixed partition approach will lose power if the affected set of locations is divided between multiple partitions rather than corresponding to a single partition. While ad-hoc partitioning methods allow partitions to vary in size, the chosen set of partitions still may not correspond to