Research Paper

Understanding the Correlations between Social Attention and Topic Trends of Scientific Publications

Xianlei Dong[1], Jian Xu[2], Ying Ding[3], Chenwei Zhang[3], Kunpeng Zhang[4] & Min Song[5†]

[1]School of Management Science and Engineering, Shandong Normal University, Jinan 250014, China
[2] School of Information Management, Sun Yat-sen University, Guangzhou 510006, China
[3]Department of Information and Library Science, Indiana University, Bloomington, IN 47405, USA
[4]Department of Information and Decision Sciences, University of Illinois at Chicago, IL 60607, USA
[5]Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seoul 120-749, Republic of Korea

Citation: Xianlei Dong, Jian Xu, Ying Ding, Chenwei Zhang, Kunpeng Zhang & Min Song (2016). Understanding the Correlations between Social Attention and Topic Trends of Scientific Publications.
Received: Jan. 18, 2016
Revised: Feb. 26, 2016
Accepted: Feb. 27, 2016

Abstract

Purpose: We propose and apply a simplified nowcasting model to understand the correlations between social attention and topic trends of scientific publications.

Design/methodology/approach: First, topics are generated from the obesity corpus by using the latent Dirichlet allocation (LDA) algorithm and time series of keyword search trends in Google Trends are obtained. We then establish the structural time series model using data from January 2004 to December 2012, and evaluate the model using data from January 2013. We employ a state-space model to separate different non-regression components in an observational time series (i.e. the tendency and the seasonality) and apply the "spike and slab prior" and stepwise regression to analyze the correlations between the regression component and the social media attention. The two parts are combined using Markov-chain Monte Carlo sampling techniques to obtain our results.

Findings: The results of our study show that (1) the number of publications on child obesity increases at a lower rate than that of diabetes publications; (2) the number of publication on a given topic may exhibit a relationship with the season or time of year; and (3) there exists a correlation between the number of publications on a given topic and its social media attention, i.e. the search frequency related to that topic as identified by Google Trends. We found that our model is also able to predict the number of publications related to a given topic.

Research limitations: First, we study a correlation rather than causality between topics' trends and social media. As a result, the relationships might not be robust, so we cannot

JDIS
Journal of Data and Information Science
Vol. 1 No. 1, 2016
pp 28–49
DOI: 10.20309/jdis.201604

† Corresponding author: Min Song (E-mail: min.song@yonsei.ac.kr).

predict the future in the long run. Second, we cannot identify the reasons or conditions that are driving obesity topics to present such tendencies and seasonal patterns, so we might need to do "field" study in the future. Third, we need to improve the efficiency of our model by finding more efficient variable selection models, because the stepwise regression method is time consuming, especially for a large number of variables.

Practical implications: This paper analyzes publication topic trends from three perspectives: tendency, seasonality, and correlation with social media attention, providing a new perspective for identifying and understanding topical themes in academic publications.

Originality/value: To the best of our knowledge, we are the first to apply the state-space model to examine the relationships between healthcare-related publications and social media to investigate the relationships between a topic's evolvement and people's search behavior in social media. This paper thus provides a new viewpoint in the correlation analysis area, and demonstrates the value of considering social media attention in the analysis of publication topic trends.

Keywords Social media; Publication topic trends; Correlation; State-space model; Variable selection; Nowcasting

1 Introduction

We live in a data-driven society where social media platforms allow users to generate content to share, communicate, and discuss their opinions with each other on various events without being at the same time and in the same place. The effect of social media on our lives has never been greater on personal as well as political and social concerns, where its impacts can be seen from local and national to global arenas. An emergence of societal and political topics thus corresponds with the traffic of social media discussions. For example, news and comments about natural disasters and political uprisings travel at breakneck speed in Twitter or Google. Social media are becoming an important channel that can also greatly influence our perception of a certain topic. In economics, social media can have a significant effect on a firm's reputation, sales, and even survival (Kietzmann et al., 2011). Altmetrics (Priem et al., 2010) use social media to estimate the early impact of publications or researchers. Dong and Bollen (2015) applied Google search-engine query data to detect consumer confidence indexes. Scott and Varian (2014) predicted weekly initial claims for unemployment and monthly retail sales using Google Trends and Google Correlate data. These evidences indicate that a range of patterns can be discovered by analyzing people's behaviors and their topically relevant online activities in social media.

In a similar vein, the proposed work seeks to understand whether there is the correlation between the major social media Google Trend and scientific topics in

academic publications. Correlation analysis identifies the degree of relationship or dependency between two types of variables. For example, the Pearson product-moment correlation coefficient (PMCC) and rank-correlation coefficient are widely used in linear relationship analysis (Kendall, 1962; Rodgers & Nicewander, 1988). PMCC and rank-correlation coefficient are not robust correlations, however, especially when there are outliers in samples; consequently, the result might be different from the truth (Wilcox, 2005). And when other things significantly influence the objective index, it becomes difficult to find a correct correlation. We therefore propose a new method to find correlations between scientific topic trends in academic publications and the social media attention they garner by means of a regression analysis between these social media and scientific topic trends. We first divide the scientific topic trends in academic publications into three parts: tendency, seasonality, and correlations with social media attention that are named the regression component. We then analyze the correlation between the regression component and the social media attention.

We approach the possibility of these correlations with the concept of "seasonality" that describes the observable patterns of topical tendency over time. Seasonality helps us understand a topic's cyclic changes by accounting for fluctuation in its own dynamics. A number of seasonality findings have been presented in the economics and engineering literature. For example, Scott and Varian (2014) proposed a nowcasting (a contraction of "now" and "forecasting") model commonly used in economics and meteorology to separate tendency, seasonality, and regression effects from economic phenomena. Analyzing topic seasonality is critical to understanding how a topic evolves in a characteristic pattern, especially for research related to topic trends and prediction. In this paper, we quantify a seasonality factor as well as overall dynamic tendency and relationships between topic evolutions and potential indicators. To this end, we choose the seasonable topic "obesity" with a particular focus on two commonly representative sub-topics, "child obesity" and "diabetes," both of which are major concerns in current health initiatives due to their increasing prevalence.

On a global scale, obesity has more than doubled since 1980 where two million children under age five were overweight or obese in 2013 (WHO, 2015). In America the large majority of American adults are obese, making it a nationwide epidemic. The condition has been implicated as a leading factor in deadly diseases such as heart disease, stroke, and diabetes. Obesity has become a focus of attention for broader audiences, including scientific researchers, social media users, medical experts, and patients themselves. Obesity-related topics, including diet, lifestyle, and child obesity, are frequently discussed in both academic publications and non-professional social media.

A recent study indicated that obese youth are likely to have a risk of cardiovascular disease (e.g. high cholesterol or high blood pressure), where among boys from age 5 to 17, 70% are at risk for cardiovascular disease (Freedman et al., 2007). Obese adolescents are more likely to have pre-diabetes conditions, where their blood glucose levels indicate a high risk of diabetes (Li et al., 2009; Centers for Disease Control and Prevention, 2011). Another study found that children with obesity as young as age two are likely to have obesity later in life (Freedman et al., 2005). Obesity at an early age may often lead to weak bones and joints, sleep apnea, and numerous social and psychological problems, such as stigmatization and poor self-esteem (Daniels et al., 2005; Dietz, 2004). It is also likely to trigger diseases later in life, including heart disease, Type 2 diabetes, stroke, osteoarthritis, and several types of cancer. These include organ cancers such as breast, colon, endometrium, esophagus, kidney, pancreas, gall bladder, thyroid, ovary, cervix, and prostate, as well as multiple myeloma and Hodgkin's lymphoma (Kushi et al., 2006). Extensive research efforts and accomplishments related to obesity have thus been conducted and published by academic scholars, which leads to a large number of digital texts. Detailed textual analyses on issues such as topic extraction, topic evolution, and topic trending are important for gaining a comprehensive understanding of obesity for better healthcare initiatives and health policy planning.

We propose a simplified nowcasting model that combines stepwise regression and Bayesian sampling to describe the relationships between obesity topics and social media. The experimental results show that the proposed state-space model can capture the impact of dynamic tendency and seasonality, and the impact of public attention from social media (e.g. Google Trends). To the best of our knowledge, we are the first to apply the state-space model to examine the relationships between healthcare-related publications and social media to investigate the relationships between a topic's evolvement and people's search behavior in social media. The proposed study investigates the nature of a topic's evolvement in three components (i.e. dynamic tendency, seasonality, and correlation with social media) and the ways in which it evolves (i.e. quantification of the above three components).

This paper is organized as follows: the data and methodology section describes the proposed model. The result section explains the experiment results for the topics of "child obesity" and "diabetes." The discussion section evaluates the performance of the model. Finally, the paper summarizes and suggests future work.

2 Data and Methodology

2.1 Data

This paper proposes to study topics involved in obesity publications, seasonality patterns behind these topics, and the relationships between the topics and social

media attention. We use academic papers and online search queries to derive topics and social attention based on two datasets: (1) academic papers about obesity in PubMed during a specific time period, and (2) Google Trends data for a specific number of search queries over the same time period.

2.1.1 Data from PubMed

For study data we downloaded obesity-related papers from PubMed for the period of January 2004 to January 2013. We use data from January 2004 to December 2012 for modeling and the January 2013 data for evaluation. Search strategies are based on the following terms (including plurals and variants) as determined by checking the Unified Medical Language System (UMLS) and consulting medical domain experts: OBESITY, OBESE, ADIPOSITY, OVERWEIGHT, EXCESS FOOD INTAKE, FEAR OF BECOMING FAT, LEPTIN, and BARIATRIC. In total, 98,063 articles (henceforth the "obesity paper dataset") are collected.

2.1.2 Data from Google Trends

Google Trends[①] is a tool offered by Google for obtaining search-query data. It takes any user search query and returns a weekly or monthly normalized time series (the maximum number of data points in the time series equals 100) of the related trend on that search query. For example, Figure 1(a) and 1(b) show the weekly search popularity of "obesity" and the monthly search popularity of "14" in the "people & society" category on Google, respectively, for the period under consideration. The overall monthly search popularity on the topic "child obesity" (i.e. the average search trend of all the queries related to this topic) is displayed in Figure 1(c). It can be clearly seen that, for all the queries related to the topic "child obesity," the overall search trend is increasing, but when it comes to certain individual indicators, both downward trends and upward trends are possible.

The data are normalized by the overall search volumes. It is scaled so that the maximum time series number equals 100, i.e. values on *y*-axis equal the search volume at *x*-axis divided by the biggest value in the whole time period. Decreases between 2004 and 2012 therefore indicate that the absolute search volume for the word "obesity" is becoming less, while the contrary is true for word "14." On average, however, the search volume for topic "child obesity" has an upward trend.

Google Trends data express some trends of search queries, i.e. the change in search volume of queries over time. Meanwhile, the change in the number of obesity-related papers over time could also be seen as a type of trend. One of our

① www.google.com/trends

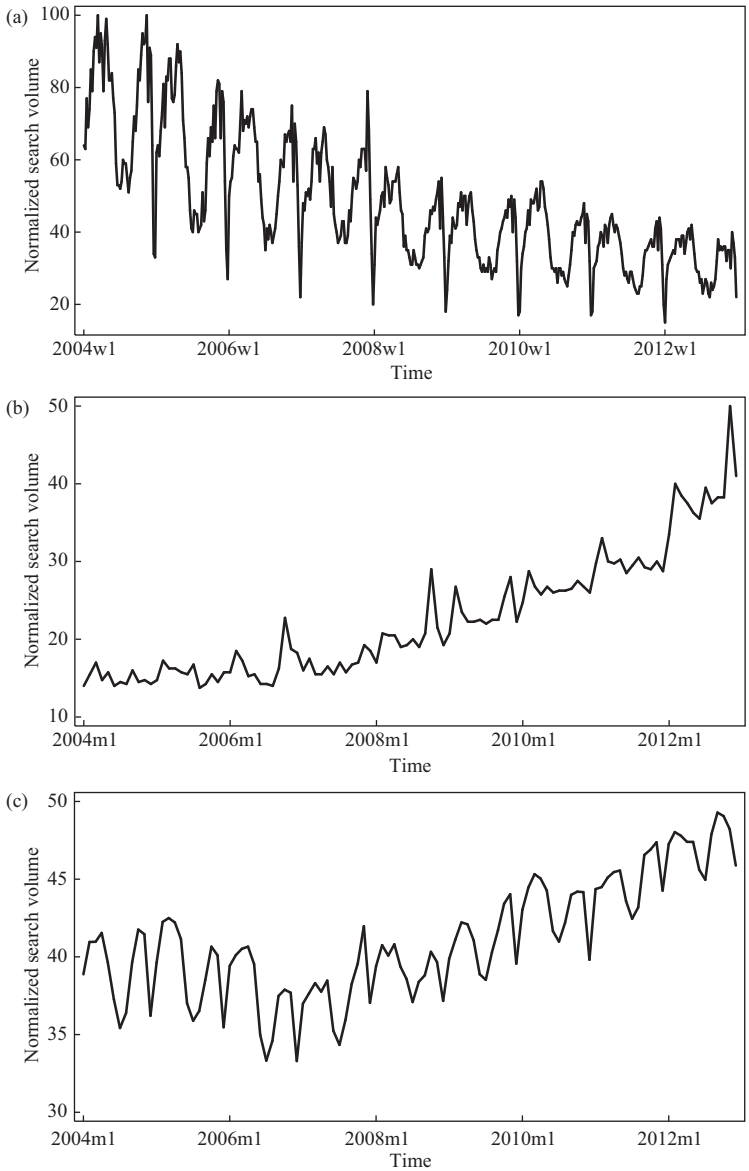

Figure 1. Google Trends graph showing (a) weekly search popularity of "obesity," (b) monthly search popularity of "14," and (c) average search trend of all the queries related to topic "child obesity."

tasks herein is to find a relationship between these two trends. The search query's context can be set to several categories, such as Arts & Entertainment, Finance, Games, and Health. In this paper, we choose seven categories where people might talk about obesity: Beauty & Fitness, Food & Drink, Health, Hobbies & Leisure, Jobs & Education, People & Society, and Sports. Categories such as Shopping or Pets & Animals are not included. PubMed data and Google Trends time-series data can be matched. Since Google Trends data can be provided weekly and PubMed data are released monthly, we convert all weekly data to monthly by taking a four-week moving average. For every selected topic discussed above, we obtain Google Trends time-series data from January 2004 to January 2013.

2.2 Methodology

The overall framework of the methodology is shown in Figure 2, including generating topics from the obesity corpus using the latent Dirichlet allocation (LDA) algorithm (Blei, Ng, & Jordan, 2003), obtaining time series of keyword search trends in Google Trends, training the structural time series model using data from January 2004 to December 2012, and evaluating the model using data from January 2013.

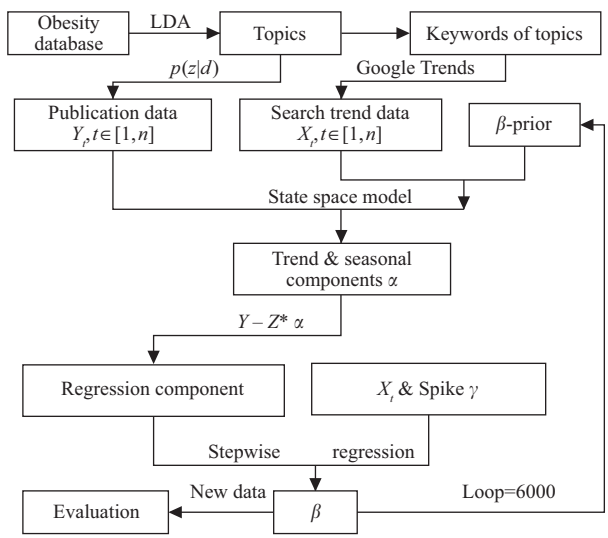

Figure 2. The overall framework of the methodology, where (1) $p(z|d)$ denotes the probability that document d belongs to topic z; (2) β denotes the keywords' effects on topics, that is, the coefficients of X; (3) spike γ can make most of the coefficients of X zeros, which ensures that the stepwise regression process will run correctly; and (4) $Y - Z^{*}\alpha$ (regression component) refers to publication data with the time-series component, where tendency and seasonal components are not included.

In this paper, we employ a state-space model to separate different non-regression components in an observational time series (i.e. the tendency component and the seasonality component) and apply the "spike and slab prior" and stepwise regression to analyze the correlations between the regression component and the social media attention. We combine the two parts using Markov-chain Monte Carlo (MCMC) sampling techniques to make the model run continuously, step by step:

1) Data preparation: Using LDA, we obtain the two topics (consisting of a number of keywords) from the obesity database (PubMed articles), and determine the probability that a particular document *d* belongs to a specific topic *z*. Based on these results, publication data could come from the probability that document *d* belongs to topic *z*. At the same time, we can get social media attention data (search volume data) referring to Google Trends.
2) State-space model: In the section Model Training, the first three sub-sections (i.e. spike and slab prior, prior of β and σ_ε^2 and prior of $\sigma_u^2, \sigma_v^2, \sigma_w^2$) represent preparatory work to realize and solve the model. When we get the results of the state-space model from the sub-section Simulate, the latent state α, we can separate the regression component from the scientific topics trends.
3) Variables selection and regression: We then use the spike and slab prior for the first step of variables selection, and use stepwise regression to finish this work. Last, we can get a correlation between the regression component and the social media attention variables, i.e. a new β (sub-section 4: estimate β and simulate σ_ε^2).
4) Repetition of Step 2 and Step 3: Markov-chain Monte Carlo (sub-section 5) is used to repeat Step 2 and Step 3 to obtain convergent results.
5) Model evaluation: Forecasting (sub-section 6) could be used to evaluate our model to some degree.

2.3 Structural Time-series Model

The state-space model is one of the most popular methods used to solve the time-series problem. Commonly used for dynamic analysis, the model provides a unified methodology for treating a wide range of problems in time-series analysis (Durbin & Koopman, 2001). State-space time-series analysis began with Kalman (1960) and has been widely applied in engineering, economics, and social sciences.

State-space methods have yielded valuable results in recent years. For example, Rueda and Rodríguez (2010), in a study estimating and forecasting fertility rates, introduced multivariate state-space models that are dynamic alternatives to logistic representations for fixed time points. Costa and Alpuim (2010) contributed to the problem of state-space model parameter estimation by proposing estimators for the

mean, the autoregressive parameters, and the noise variances. Al-Anaswah and Wilfling (2011) used a state-space model with Markov-switching to detect speculative bubbles in stock-price data, and found that in the stock markets considered, their identification procedure correctly detects most speculative periods that have been classified as such by economic historians. Unnikrishnan (2012) made a prediction of magnetic sub-storms using a state-space model, generating outputs for storm events that reasonably reproduce the observed values, which demonstrates its prediction capability. Dong et al. (2014) focused on developing flexible and explicitly multivariate state-space models for network flow rate and mean-speed predictions. Using two-minute measurements from an urban freeway network, they provided practical guidance for selecting the most appropriate models for congested and non-congested conditions. Ghosh et al. (2014) introduced Bayesian inference in nonparametric dynamic state-space models, and illustrated their methods with simulated datasets, using the Markov-chain Monte Carlo (MCMC) approach for studying posterior distributions of interest.

The unique contribution of the state-space model is its modularity. It can capture the relationship between a dependent variable and each individual independent variable. In this paper, we apply a state-space model to separate the tendency, seasonality, and regression component from the publication trends under a scientific topic. Each of these three components can be modeled separately (Draper & Smith, 1998). Tendency of a topic reflect quantitative changes (increase or decrease) in publication volume over time, which may result from the existence of more authors and the development of technology over time. Seasonality is a time-series-based pattern that is predictable and can be repeated at different periods within a year. Seasonality tells us the periodic change in publication numbers of a topic over a year, which may result from the publication cycle of related journals and conferences. The regression component is something that correlates with the public's social attention that is used to determine the popularity of a topic based on the frequency of query keywords. Due to its modularity and capability in separating and evaluating regression components individually, we apply a state-space model to study impacts of the tendency, seasonality, and regression component on the number of publications.

Let y_t denote observation t in a real-valued time series. The state-space model can be formulated in many ways, such as a linear Gaussian (i.e. obeying normal distribution):

$$\begin{cases} y_t = Z_t\alpha_t + \varepsilon_t, \varepsilon_t \sim N\left(0, H_t\right) \\ \alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \eta_t \sim N\left(0, Q_t\right) \end{cases}, t = 1, 2, \cdots, n. \quad (1)$$

There are two equations in this model, where the first one is known as the observation equation because it links the observed variable with the unobserved

latent state variable, and the second one is known as a transition equation because it defines an evolvement process for the unobserved variable. Where y_t represents the current observation variable, α_t represents the latent state variable, Z_t and T_t are the coefficient matrixes of observation equation and transition equation, respectively, and ϵ_t and $R_t\eta_t$ are disturbances of observation equation and state equation, respectively. R_t could assure Q_t is a full rank variance matrix.

Yet because we want the model to capture effects of the three aspects of tendency, seasonality, and regression component, the model could be rewritten as:

$$\begin{cases} y_t = \mu_t + \tau_t + x_t\beta + \varepsilon_t \\ \mu_{t+1} = \mu_t + \delta_t + u_t \\ \delta_{t+1} = \delta_t + v_t \\ \tau_{t+1} = -\sum_{j=1}^{s-1} \tau_{t+1-j} + w_t \end{cases}, \tag{2}$$

where μ_t, δ_t, and τ_t represent the current level of the tendency, the current "slope" of the tendency, and the seasonal component, respectively, and $x_t\beta$ represents the regression component. Generally, the seasonal component could be set into a set of s dummy variables with dynamic coefficients constrained to have zero expectation over a whole cycle.

Following the methods discussed in Scott and Varian (2014), we assign the regression effects $x_t\beta$ to Z_t rather than to α_t, which would increase the dimension of α_t only by one. The computational complexity of the Kalman Filter (Kalman, 1960) is linearly related to the length of the data, and quadratically related to the size of the state. The parameters in our model can therefore be written as:

$$\alpha_t = \left(1, \mu_t, \delta_t, \tau_t, \tau_{t-1}, \cdots, \tau_{t-s+2}\right)^T, \tag{2.1}$$

$$Z_t = \left(x_t\beta, 1, 0, 1, 0, \cdots, 0\right), \tag{2.2}$$

$$T_t = diag\left(T_\mu, T_\tau\right), \tag{2.3}$$

$$\eta_t = \left(u_t, v_t, w_t\right), \tag{2.4}$$

$$R_t = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.5}$$

$$H_t = \sigma_\varepsilon^2, \tag{2.6}$$

$$Q_t = diag\left(\sigma_u^2, \sigma_v^2, \sigma_w^2\right), \tag{2.7}$$

$$\text{in which, } T_\mu = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{2.8}$$

$$T_\tau = \begin{bmatrix} -1 & -1 & \cdots & -1 & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \tag{2.9}$$

where $diag\left(\sigma_u^2, \sigma_v^2, \sigma_w^2\right)$ refers to the diagonal elements matching these parameters, while the other elements in the matrix are zeros.

2.4 Model Training

The training of the state-space model involves variable selection and sampling techniques. Here we use the spike and slab prior and stepwise regression (Draper & Smith, 1998) for variable selection, and the Koopman smoother and MCMC methods (Durbin & Koopman, 2001) for simulation and sampling.

2.4.1 Spike and Slab Prior

Since there are a large number of variables in regression effects, and while the observation data are typically brief, we use the spike and slab prior on the regression coefficients to indicate sparsity. The spike and slab prior can make some of the coefficients of regression variables zeros. Let $\gamma_k = 1$ if $\beta_k \neq 0$, and $\gamma_k = 0$ if $\beta_k = 0$. β_γ denotes the subset of elements where $\beta_k \neq 0$. Assume that the dimension of each variable in X is K, where X represents the regression variables and K is the number of variables in X, that is, X is a matrix with n rows and K columns.

It is very common to see a Bernoulli distribution (0 – 1 distribution) as the spike and slab prior. From $\gamma\sim$ Bernoulli prior, we can then draw a $\gamma_{K\times 1}$ as such:

$$\gamma \sim \prod_{k=1}^{K} \pi_k^{\gamma_k} \left(1-\pi_k\right)^{1-\gamma_k}. \tag{3}$$

But the initial value of π_k comes from practice, even though it could be set according to some "rules." For example, we can first think about a number of non-zero variables in the regression part such as m, and then take $\pi_k = m/K$.

2.4.2 Prior of β and σ_ε^2

This assumes that the conditional prior $\sigma_\varepsilon^2|\gamma \sim IG\left(\frac{v}{2}, \frac{ss}{2}\right)$ in which $v/2$ determines the shape and $ss/2$ determines the scale of the distribution. Inverse Gamma Distribution (*IG* Distribution) is chosen due to its property that when it is set to be the prior of σ^2 after obtaining a sample $u_{1:n}$ from independent Gauss distribution $N(0, \sigma^2)$, the posterior of σ^2 will be expressed in the following form:

$$\sigma^2 \sim IG\left(\frac{c+n}{2}, \frac{r+\sum_{i=1}^{n}\mu_i^2}{2\sigma^2}\right) \text{ (Poirier, 1995).} \tag{4}$$

Initial values of v and ss are set using Equation (5):

$$ss/v = (1-R^2)s_y^2, \tag{5}$$

where s_y^2 is the marginal standard deviation of the response (the standard deviation of observation variable y_t), and the expected R^2 is from the regression equation. In this paper, we use 0.5 and 0.01 for ss and v, respectively. Then we can obtain s_y^2 from observation $y_{1:n}$, compute the value of ss, and obtain the initial value of σ_ε^2. However, the ss can be chosen at will, as long as the algorithm can be successfully run, and any ss can be used to represent ignorance (George & McCulloch, 1993).

According to Zellner's *g*-prior (Liang et al., 2008), we can define an Ω^{-1} as follows: Let X denote the design matrix, i.e. $X = (x_1, x_1, ..., x_n)^T$, where x_t denotes an observation at time t. A full-rank matrix could then be set as Equation (6):

$$\Omega^{-1} = \frac{\kappa}{2n}\left[X^T X + diag\left(X^T X\right)\right], \tag{6}$$

where $diag(X^T X)$ is the diagonal matrix with diagonal elements matching those of $X^T X$, while other elements are zeros. We set $\kappa = 1$ in our paper, where n is the number of real-value time series ($y_{1:n}$). For the symmetric matrix Ω^{-1}, let Ω_γ^{-1} denote the rows and columns of Ω^{-1} related to $\gamma_\kappa = 1$.

Then the conditional prior of $\beta_\gamma|\sigma_\varepsilon^2, \gamma$ is:

$$\beta_\gamma|\sigma_\varepsilon^2, \gamma \sim N\left(b_\gamma, \sigma_\varepsilon^2\left(\Omega_\gamma^{-1}\right)^{-1}\right), \tag{7}$$

where $b_\gamma = 0$ (it is common to set b_γ as a zero matrix). Then we can get β_γ (the regression variables' coefficients that are not 0), i.e. after adding the zeros in, we get a β.

2.4.3 Prior of $\sigma_u^2, \sigma_v^2, \sigma_w^2$

Let θ denote the set of model parameters other than β and σ_ε^2 i.e. $\theta = \left(\sigma_u^2, \sigma_v^2, \sigma_w^2\right)$. An Inverse Gamma full conditional distribution is used to gain the prior value of θ, i.e.:

$$\sigma_u^2 \sim IG\left(\frac{c_1}{2}, \frac{r_1}{2}\right),$$
$$\sigma_v^2 \sim IG\left(\frac{c_2}{2}, \frac{r_2}{2}\right), \quad (8)$$
$$\sigma_w^2 \sim IG\left(\frac{c_3}{2}, \frac{r_3}{2}\right).$$

The initial values of c and r could be 1 and 80. We then derive a θ from the IG distribution.

2.4.4 Simulate the Latent State $\alpha(\tilde{\alpha})$

There are many methods available for solving the state-space model, such as the Kalman filter (Andrew, 1989; Kalman, 1960) and Bayesian computation methods (De Jong & Shephard, 1995; McCausland, Miller, & Pelletier, 2011). As β, σ_ε^2 and θ are known, we can simulate the latent state $\alpha = \tilde{\alpha}$ using a Koopman smoother (Durbin & Koopman, 2001). A Koopman smoother is one of the primary tools for working with state-space models. It simulates α from $p\left(\alpha | y_{1:n}\right)$ under the assumption that a_1 and P_1 are known, where $\alpha_1 \sim N(a_1, P_1)$. It then modifies the initial conditions as P_1 has infinite variance.

2.4.5 Simulate θ

Using $\tilde{\alpha}$, $X\beta$ and the observation data $y_{1:n}$, we can get the error terms $u_{1:n}$, $v_{1:n}$, $w_{1:n}$ from Equation (1). The corresponding post prior of θ would thus be expressed in the following forms:

$$\sigma_u^2 \sim IG\left(\frac{c_1 + n}{2}, \frac{r_1 + \sum_{i=1}^{n} u_i^2}{2\sigma_u^2}\right),$$
$$\sigma_v^2 \sim IG\left(\frac{c_2 + n}{2}, \frac{r_2 + \sum_{i=1}^{n} v_i^2}{2\sigma_v^2}\right), \quad (9)$$
$$\sigma_w^2 \sim IG\left(\frac{c_3 + n}{2}, \frac{r_3 + \sum_{i=1}^{n} w_i^2}{2\sigma_w^2}\right).$$

2.4.6 Estimate β and Simulate σ_ε^2

We can select the most important independent variables from X (K dimension) and calculate the correlation between regression component and social media attention data using the following steps:

a) Denote Z_t^* as the Z_t with $x_t\beta = 0$; let $y_t^* = y_t - Z_t^*\alpha, t = 1,2,\ldots,n$. Then let $y^* = y_{1:n}^*$, i.e. y^* is y with the time-series component subtracted out (regression component);
b) Then, based on the γ we obtain, we can get an X_γ, where X_γ denotes the rows of X corresponding to $\gamma_k = 1$;
c) Estimate a β_γ using y^* and X_γ using the method of stepwise regression; and
d) Get a β using the β_γ and γ.

Within Equation (1), Z_t, α_t having been simulated, we can now compute the error ε and update the σ_ε^2 using Equation (10):

$$\sigma_\varepsilon^2|\gamma, y^* \sim IG\left(\frac{v+n}{2}, \frac{ss+|\, y^* - X\beta\,|^2}{2\sigma_\varepsilon^2}\right), \tag{10}$$

where $\left|y^* - X\beta\right|^2 = \sum_{i=1}^{n}\left(y_i^* - x_i\beta\right)^2$. After the above computations, renew the γ from its Bernoulli prior.

2.4.7 Markov-chain Monte Carlo

a) Simulate α from $p(\alpha|y, \theta, \beta, \sigma_\varepsilon^2)$ based on the known y, θ, β and σ_ε^2 (prior ones or posterior ones from the following steps);
b) Simulate θ from $p(\theta|y, \alpha, \beta, \sigma_\varepsilon^2)$ based on the known y, β, σ_ε^2 and the simulated α;
c) Estimate β from $p(\beta|y, \alpha, \theta, \sigma_\varepsilon^2)$ using stepwise regression, based on the known y, X and the simulated α; and
d) Simulate σ_ε^2 from $p(\sigma_\varepsilon^2|y, \alpha, \theta, \beta)$ based on the known y, the simulated α and the estimated β.

Repeated cycling through the four steps above generates a sequence of draws $\phi^{(1)}$, $\phi^{(2)}$, $\phi^{(3)}$, … (say, 6,000 times). These $\phi^{(i)}$ ($i \geq M$, let M be a sufficiently large integer) could be seen as generated using MCMC with stationary distribution $p(\phi|y)$.

2.4.8 Forecasting

This model also has prediction ability. For every ϕ, we can compute an α_{n+1} using Equation (1). Then we can obtain y_{n+1} after estimating $x_{n+1}\beta$ if a Z_{n+1} is known. The forecasting work can be done using the mean:

$$\tilde{y}_{n+1} = \frac{\sum_{i=1}^{M} y_{n+1}^{(i)}}{K}. \tag{11}$$

Research Paper

3 Results

In this paper we investigate impacts of the three components defined in our state model on the number of publications from PubMed data: tendency, seasonality, and influence of the regression component correlated with Google Trends data. Each PubMed article related to obesity consists of fields of PubMed ID, title, abstract, date, authors, and others. LDA is applied to obtain latent topics from this textual information (Blei, Griffiths, & Jordan, 2010). The number of topics (k) is set to be 50. Two representative sub-topics ("child obesity" and "diabetes") are chosen for analysis.

The number of publications ($N(m)^t$) related to a specific topic (t:$1 \le T \le k$) within the m^{th} month can be calculated from the following two steps: (1) topic distribution $\left(p_i^1, p_i^2, \cdots, p_i^k\right)$ for each paper/document (d_i) can be obtained after applying LDA to papers published in a given month, where the probability (p_i^j) represents the percentage of j^{th} topic in that document; and (2) $N(m)^t$ can be obtained by summing the p_i^t of all documents, $N(m)^t = \sum_{1 \le i \le n} p_i^t$.

In the results of LDA topics modeling, two topics are highly related to diabetes. We thus combine them into one topic based on the additive law of probability, i.e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. In our LDA model, $P(A \cap B) = 0$, because for simplicity, we assume that all topics are independent. Figure 3 shows the monthly change in the number of publications on the topics "child obesity" and "diabetes," respectively.

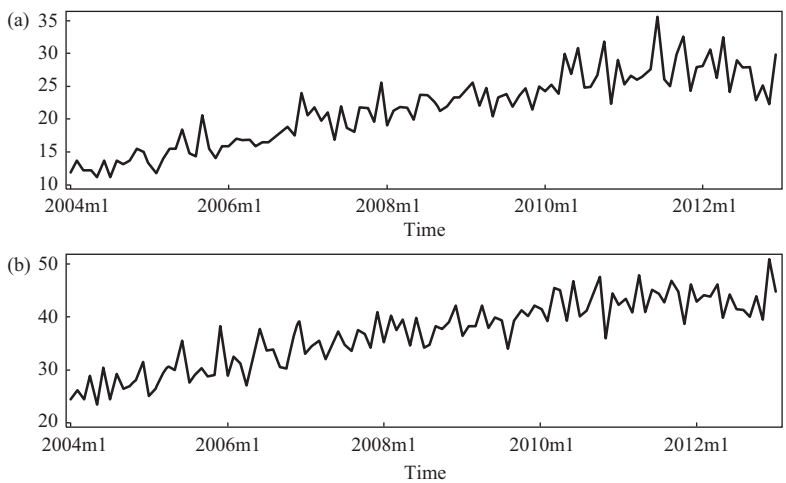

Figure 3. The monthly number of publications on (a) "child obesity" and (b) "diabetes" over time.

Another output from LDA is the topic-term distribution. For each topic, every word has a corresponding probability, indicating its likelihood belonging to that topic. For topics "child obesity" and "diabetes," we choose 50 and 100 keywords with the highest probabilities and use them as search-query keywords in Google Trends. Each word is searched under seven different categories defined by Google, to obtain seven time-series-based trends. If the number of queries using one of these words is too small to have an explicit trend, the word under that category is dropped. We eventually find 344 and 616 search volume trends for topics "child obesity" and "diabetes," respectively.

3.1 Tendency

We define two time-series-based measures of the publication topic trends (i.e. tendency and seasonality) here for the topics "child obesity" and "diabetes." The tendency of a topic, the first component of the publication topic trends, is the change in the number of publications for that topic over time. It represents the overall popularity of that topic at different time periods. Changes in these tendencies could result from many aspects, such as an increase in the scholars in this field, related technology under rapid development, and more research findings emerging over time. Figure 4(a) and 4(b) show the tendencies for the two chosen topics. The increasing trends indicate that the number of publications related to these two topics generally increased from 2004 to 2014, except for some fluctuations between 2004 and 2006. It also indicates that a growing amount of scientific research is being directed to these two topics, as the scholars in this field are publishing more research.

It is still difficult to detect the rate of change by viewing only these tendencies, however. We use "growth rate" to capture the changing rate. Figure 4(c) and 4(d) show growth rates for the two topic tendencies, where the average growth rate for "child obesity" is 0.158 and 0.183 for "diabetes." This means that the number of publications on "child obesity" increases more slowly than the number for "diabetes." While research on "child obesity" is very popular, research on "diabetes" has become more popular in recent years.

3.2 Seasonality

Tendency and growth rate can help us explain the overall characteristics of topic evolution, such as their increase (or decrease) or speed of increase (or decrease) in popularity over different time periods. However, they do not make it easy to answer the following questions: what are major factors leading to these fluctuations, and are there any patterns behind these changing curves? To resolve these issues, considering the publication of related periodicals and the convening of related meetings are determined to some degree. The truth would result in a common

Research Paper

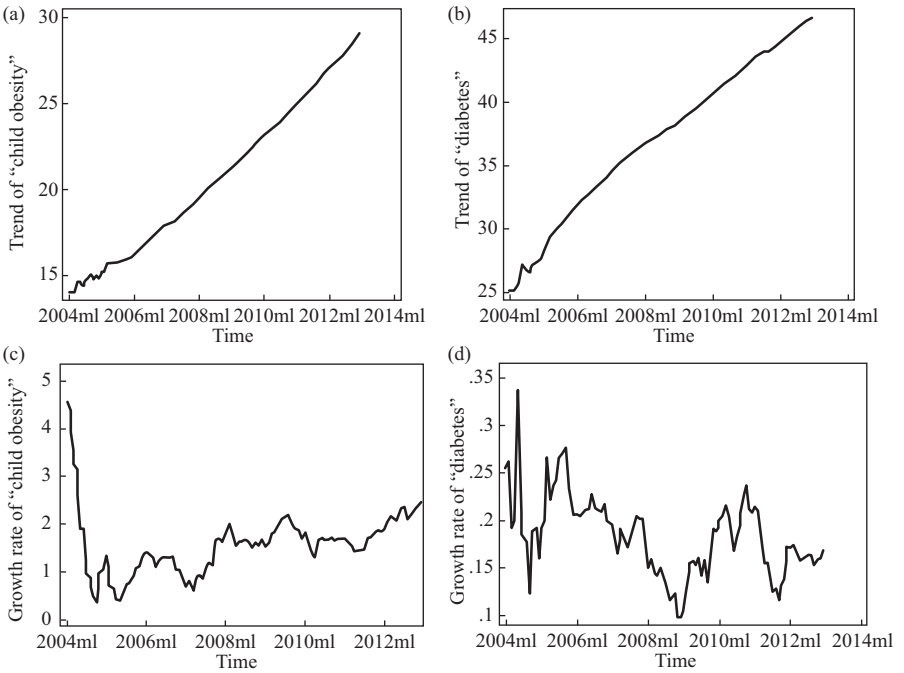

Figure 4. Trends of topics (a) "child obesity" and (b) "diabetes." The *x*-axis represents time from January 2004 to January 2013; the *y*-axis represents the number of publications within a month on "child obesity" and "diabetes," respectively. Growth rate of topics (c) "child obesity" and (d) "diabetes." The *x*-axis represents time from January 2004 to January 2013; the *y*-axis represents the growth rate of publications within a month on "child obesity" and "diabetes," respectively.

phenomenon: the publication numbers on the two chosen topics increase in some months and decrease in other months, where the same is true every year, i.e. some of these changes are repeated periodically. This phenomenon, defined as seasonality in this paper, is seen in many domains. For example, sales of down jackets increase in fall and winter and decrease in spring and summer, data that have remained consistent from year to year for obvious reasons. To capture the seasonal effect on topics among publications (which are not obvious) we consider seasonality as a separate component in our model. As shown in Figure 5, there are more publications on "child obesity" published in June, September, and December than in January, May, and November. Similarly, more publications on "diabetes" are published in April, June, and December than in January, May, July, and November. The number of publications on a topic in a time period might indicate the degree of interest in that topic for a given period.

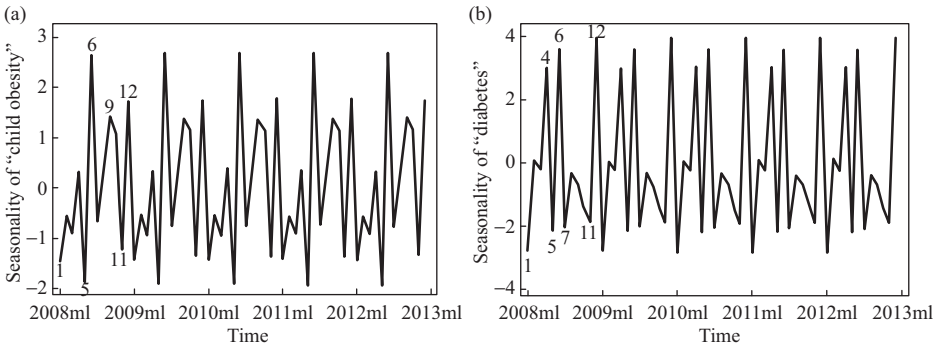

Figure 5. Seasonal effect for topics (a) "child obesity" and (b) "diabetes" from January 2008 to January 2013.

3.3 The Regression Component Correlating with Social Media Attention

The extent of interest in obesity and diabetes can be observed in two ways: the number of scientific publications in related journals and conferences, and the frequency of related query terms in Google Trends used by the public. In our model, we have a regression component that describes the correlation between Google Trends data and the number of publications. The regression components for "child obesity" are mostly positive, while negative or close to zero for "diabetes," as shown in Figure 6. A positive x means that the search query data from Google Trends are directly proportional to the number of scientific publications. In addition, our model can discover the impact of individual keywords from each category of Google Trends on the number of publications. The magnitude of this impact is modeled by the parameter β in our model. There are 344 values for "child obesity" and 616 for "diabetes." Yet in spite of this outcome, there exists a weak impact for each individual keyword, where the correlation is strong when all the keywords' attention trends are considered.

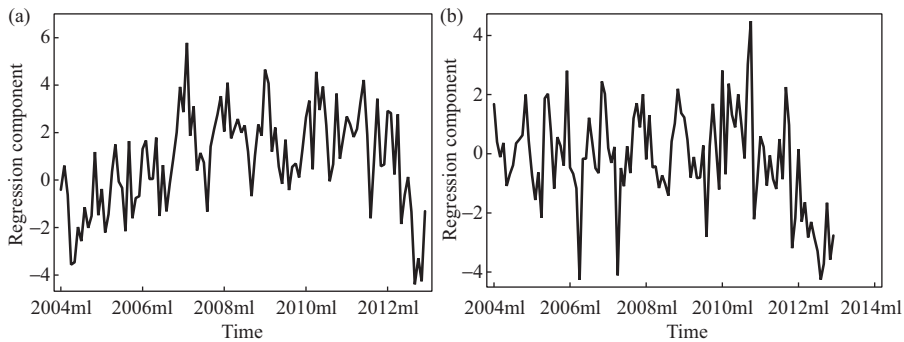

Figure 6. Regression components for obesity topics (a) "child obesity" and (b) "diabetes" over time.

4 Conclusion and Future Work

In this paper, we propose a state-space model to capture individual factors of tendency, seasonality, and regression component (correlated with social media attention in Google Trends) in the number of publications for a topic, to help us comprehensively understand the topic, including what the topic and its evolution is, how it evolves, and its relationships with social media. We choose two commonly representative sub-topics of "child obesity" and "diabetes" as cases to demonstrate our findings. We use stepwise regression for variable selection, combined with MCMC to train our model. The experimental results show that (1) the number of publications on "child obesity" increases at a slower rate than that of "diabetes" publications, which indicates that the research on "diabetes" is becoming more popular in recent years; (2) different topics exhibit different seasonal patterns in terms of the number of publications on such topics. There are more publications on "child obesity" published in June, September, and December than in January, May, and November. Similarly, more publications on "diabetes" are published in April, June, and December than in January, May, July, and November; and (3) there exists a relationship between the number of publications on a given topic and the search frequency of terms related to that topic on Google Trends.

In spite of the promise of this novel approach to capturing factors of topics, there are a number of shortcomings that should be addressed in future work. For example, we study a correlation rather than causality between topics' trends and social media. As a result, the relationships might not be robust, so we cannot predict the future in the long run. Also, we cannot identify the reasons or conditions that are driving obesity topics to present such tendencies and seasonal patterns. This limitation may be caused by government policy, climate change, or even new technology. To find the causal relationships, we need to do "field" study in the future. In addition, we need to improve the efficiency of our model by finding more efficient variable selection models, such as Bayesian inference, because the stepwise regression method is time consuming, especially for a large number of variables. Furthermore, we will try to model topics using related indicators that express social attention in other forms of social or traditional media, such as blog data or newspaper data, so we can make our model more accurate by triangulating or cross-validating results. We also want to incorporate sub-model techniques that could transform into our algorithm higher dimensions of big data into a certain number of relatively smaller data subsets, as described by Zhou et al. (2013), which might make our model more efficient.

Acknowledgements

The authors thank Beibei Hu, Hui Yang, Vincent Malic, and Sen Wu for their kind help and comments on this manuscript. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-2012S1A3A2033291) and by the Yonsei University Future-leading Research Initiative of 2014.

Author Contributions

X.L. Dong (sddongxianlei@163.com) conceived and designed the model, drafted the manuscript, played an important role in interpreting the results. J. Xu (xujianonline@qq.com) acquired data and wrote the algorithm of the model. Y. Ding (dingying@indiana.edu) contributed significantly to analysis with constructive discussions and manuscript preparation, approved the final version. C.W. Zhang (zhang334@umail.iu.edu) explained the data from PubMed. K.P. Zhang (kpzhangs@gmail.com) drafted and revised the manuscript. M. Song (min.song@yonsei.ac.kr, corresponding author) revised the manuscript and approved the final version.

References

Al-Anaswah, N., & Wilfling, B. (2011). Identification of speculative bubbles using state-space models with Markov-switching. Journal of Banking & Finance, 35(5), 1073–1086.

Andrew, H.C. (1989). Forecasting, structural time series models and the Kalman filter. Cambridge, UK: Cambridge University Press.

Blei, D.M., Griffiths, T.L., & Jordan, M.I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM, 57(2), article no. 7.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Centers for Disease Control and Prevention. (2011). National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, 2011.

Costa, M., & Alpuim, T. (2010). Parameter estimation of state space models for univariate observations. Journal of Statistical Planning and Inference, 140(7), 1889-1902.

Daniels, S.R., Arnett, D.K., Eckel, R.H., Gidding, S.S., Hayman, L.L., Kumanyika, S., . . . Williams, C.L. (2005). Overweight in children and adolescents pathophysiology, consequences, prevention, and treatment. Circulation, 111(15), 1999-2012.

De Jong, P., & Shephard, N. (1995). The simulation smoother for time series models. Biometrika, 82(2), 339–350.

Dietz, W.H. (2004). Overweight in childhood and adolescence. New England Journal of Medicine, 350(9), 855–856.

Dong, C., Shao, C., Richards, S.H., & Han, L.D. (2014). Flow rate and time mean speed predictions for the urban freeway network using state space models. Transportation Research Part C: Emerging Technologies, 43, 20–32.

Dong, X., & Bollen, J. (2015). Computational models of consumer confidence from large-scale online attention data: Crowd-sourcing econometrics. PLOS One, 10(3): e0120039.

Draper, N.R, & Smith, H. (1998). Applied regression analysis (3rd ed.). New York: John Wiley & Sons.

Durbin, J., & Koopman, S.J. (2001). Time series analysis by state space methods (2nd ed.). Oxford, UK: Oxford University Press.

Freedman, D.S., Khan, L.K., Serdula, M.K., Dietz, W.H., Srinivasan, S.R., & Berenson, G.S. (2005). The relation of childhood BMI to adult adiposity: The Bogalusa heart study. Pediatrics, 115(1), 22–27

Freedman, D.S., Mei, Z., Srinivasan, S.R., Berenson, G.S., & Dietz, W.H. (2007). Cardiovascular risk factors and excess adiposity among overweight children and adolescents: The Bogalusa heart study. Journal of Pediatrics, 150(1), 12–17.

George, E.I., & McCulloch, R.E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423), 881–889.

Ghosh, A., Mukhopadhyay, S., Roy, S., & Bhattacharya, S. (2014). Bayesian inference in nonparametric dynamic state-space models. Statistical Methodology, 21, 35–48.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. Journal of Fluids Engineering, 82(1), 35–45.

Kendall, M.G. (1962). Rank correlation methods (3rd ed.). New York: Hafner Publishing.

Kietzmann, J.H., Hermkens, K., McCarthy, I.P., & Silvestre, B.S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons, 54(3), 241–251.

Kushi, L.H., Byers, T., Doyle, C., Bandera, E.V., McCullough, M., Gansler, T., . . . Thun, M.J. (2006). American Cancer Society guidelines on nutrition and physical activity for cancer prevention: Reducing the risk of cancer with healthy food choices and physical activity. A Cancer Journal for Clinicians, 56(5), 254–281.

Li, C., Ford, E.S., Zhao, G., & Mokdad, A.H. (2009). Prevalence of pre-diabetes and its association with clustering of cardiometabolic risk factors and hyperinsulinemia among US adolescents: National Health and Nutrition Examination Survey 2005–2006. Diabetes Care, 32(2), 342–347.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., & Berger, J.O. (2008). Mixtures of *g* priors for Bayesian variable selection. Journal of the American Statistical Association, 103(481), 410–423.

McCausland, W.J., Miller, S., & Pelletier,D. (2011). Simulation smoothing for state—space models: A computational efficiency analysis. Computational Statistics & Data Analysis, 55(1), 199–212.

Poirier, D.J. (1995). Intermediate statistics and econometrics: A comparative approach. Cambridge, MA: MIT Press.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved from http://altmetrics.org/manifesto/.

Rodgers, J.L., & Nicewander, W.A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42, 59–66.

Rueda, C., & Rodríguez, P. (2010). State space models for estimating and forecasting fertility. International Journal of Forecasting, 26(4), 712–724.

Scott, S.L., & Varian, H.R. (2014). Predicting the present with Bayesian structural time series. International Journal of Mathematical Modelling and Numerical Optimisation, 5(1–2), 4–23.

Unnikrishnan, K. (2012). Prediction of magnetic substorms using a state space model. Journal of Atmospheric and Solar-Terrestrial Physics, 75, 22–30.

World Health Organization (WHO). (2015). Obesity and overweight. Fact Sheet No. 311. Retrieved from http://www.who.int/mediacentre/factsheets/fs311/en/.

Wilcox, R.R. (2005). Introduction to robust estimation and hypothesis testing (3rd ed.). Waltham, MA: Academic Press.

Zhou, J., Hu, L., Wang, F., Lu, H.H., & Zhao, K. (2013). An efficient multidimensional fusion algorithm for IoT data based on partitioning. Tsinghua Science and Technology, 18(4): 369–378.