Time-Series Adaptive Estimation of Vaccination Uptake Using Web Search Queries
Dalum Hansen, Niels; Mølbak, Kåre; Cox, Ingemar Johansson; Lioma, Christina

Published in:
Proceedings of the 26th International Conference on World Wide Web Companion

DOI:
10.1145/3041021.3054251

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Time-Series Adaptive Estimation of Vaccination Uptake Using Web Search Queries

Niels Dalum Hansen
University of Copenhagen/IBM
nhansen@di.ku.dk

Ingemar J. Cox
University of Copenhagen
ingemar.cox@di.ku.dk

Kåre Mølbak
Statens Serum Institut
KRM@ssi.dk

Christina Lioma
University of Copenhagen
c.lioma@di.ku.dk

ABSTRACT

Estimating vaccination uptake is an integral part of ensuring public health. It was recently shown that vaccination uptake can be estimated automatically from web data, instead of slowly collected clinical records or population surveys [2]. All prior work in this area assumes that features of vaccination uptake collected from the web are temporally regular. We present the first ever method to remove this assumption from vaccination uptake estimation: our method dynamically adapts to temporal fluctuations in time series web data used to estimate vaccination uptake. We show our method to outperform the state of the art compared to competitive baselines that use not only web data but also curated clinical data. This performance improvement is more pronounced for vaccines whose uptake has been irregular due to negative media attention (HPV-1 and HPV-2), problems in vaccine supply (DiTeKiPol), and targeted at children of 12 years old (whose vaccination is more irregular compared to younger children).

1. INTRODUCTION AND RELATED WORK

Vaccination programs are an efficient and cost effective method to improve public health. With sufficiently many people vaccinated the population gains herd immunity, meaning the disease cannot spread. Timely actions to avoid drops in vaccination coverage are therefore of great importance. Many countries have no registries of timely vaccination uptake information, but rely for example on yearly surveys. In such countries estimations of near real-time vaccination uptake based solely on data and search frequency data, and (ii) when the time series data to be estimated are assumed to be unavailable (near real-time). To address these challenges, we combine AA with regression trees, motivated by recent research showing that random forests outperform ARIMA models on avian influenza prediction [4]. A random forest, i.e. an ensemble of decision trees, is well suited for our problem since it is easy to extend to multiple data sources.

Our method works as follows: We initially generate a set of regression trees. For each time step the ensemble of regression trees is retrained based on the initial set of trees and a weighted sum is used to make the estimation. AA is used to continuously update the weights of each tree. Each regression tree is trained based on a set of features and training samples. For each tree a feature set is drawn with replacement from the complete feature set. Training samples are selected based on time-relative indices, where index 0 corresponds to the current time step. The indices are uniformly drawn with replacement from the interval [0 : s], where s is a window size. We use trees with different window sizes to account for stationarity and non-stationarity of the signal.

Our adaptive vaccination estimation algorithm is shown in Algorithm 1, where η is the learning rate, RT a set of N regression trees, \( i \) the amount of initial training data and \( y \) the vaccination uptake.

3. EVALUATION

To facilitate direct comparison, we evaluate our method on the same data as [2]: monthly vaccination uptake of all official
Algorithm 1: Adaptive time series estimation

Require: RT, $\eta$, $i$
1: $W \leftarrow$ list with weights, initialize to be uniform
2: $X \leftarrow$ list with the first $i$ training samples
3: $Y \leftarrow$ list with the first $i$ observations of $y$
4: $\hat{Y} \leftarrow$ empty list of estimations
5: $t \leftarrow$ current time step, starting at $i+1$
6: while True do
7: \hspace{1em} $x_t \leftarrow$ receive new observation from data stream
8: \hspace{1em} for $n = 0 \to N$ do
9: \hspace{2em} Train RT[$n$] using $X$ and $Y$
10: \hspace{2em} $\hat{Y}_{temp}[n] \leftarrow$ estimation of RT[$n$] given $x_t$
11: \hspace{2em} end for
12: \hspace{1em} $Y[t] \leftarrow \sum_{n=0}^{N} W[n] \cdot \hat{Y}_{temp}[n]$
13: \hspace{1em} $\hat{y}[t] \leftarrow$ observed $y$ at time $t$
14: \hspace{1em} for $n = 0 \to N$ do
15: \hspace{2em} $W[n] \leftarrow W[n] \cdot \exp(-\eta (\hat{Y}_{temp}[n] - \hat{y}[t])^2)$
16: \hspace{2em} end for
17: \hspace{1em} $W \leftarrow$ normalize $W$
18: \hspace{1em} $X[t] \leftarrow x_t$
19: \hspace{1em} $t \leftarrow t + 1$
20: end while

The initial number of training samples, $i$, is set to 24. All algorithms are evaluated in a leave-one-out fashion, where all data prior to the data point being estimated is used for training. For our algorithm (ATSE) a parameter search is performed by randomly sampling from the following intervals: Window size interval 1-46, number of features derived from vaccination data 0-45, number of regression trees 500-10000, $\eta$ between 0.001-0.25.

Table 1 displays the root mean squared error (RMSE) between the estimated vaccination uptake and the real vaccination uptake for all methods. Our method yields the overall best performance for all methods. Our method combines an Aggregation Algorithm (AA) to automatically reduce estimation error in changing environments with regression trees. We applied our method to estimate vaccination uptake in all official Danish children vaccines, following [2], and showed that our approach overall outperformed strong baselines that assumed data stationarity. Lastly, in recent years there have been noted irregularities in their usage, such as HPV-1, HPV-2, MMR-2(12) and DiTeKiPol. Our method was particularly strong estimating uptake for vaccines with known irregularities in their usage, such as HPV-1, HPV-2, MMR-2(12) and DiTeKiPol.

4. CONCLUSION

We presented an automatic method for near real time estimation of health events using web search query data. Our method combines an Aggregation Algorithm (AA) to automatically reduce estimation error in changing environments with regression trees. We applied our method to estimate vaccination uptake in all official Danish children vaccines, following [2], and showed that our approach overall outperformed strong baselines that assumed data to be temporally regular. Our method was particularly strong estimating uptake for vaccines with known irregularities in their usage, such as HPV-1, HPV-2, MMR-2(12) and DiTeKiPol.

This work confirms recent findings that vaccination uptake can be automatically estimated only from web data, and further extends this area by accounting for irregular uptake patterns.

Funded by IBM and the Danish Agency for Science & Higher Education.

5. REFERENCES