NetSciX2026: International School and Conference on Network Science
February 17^{th}-20^{th}, 2026 - Auckland, New Zealand

A Forecast Model for COVID-19 Spread Trends Using Blog and GPS Data from Smartphones

Keywords: COVID-19 forecasting, GPS mobility and blog data, Trend decomposition and variable selection, Explanatory variable network structure

Introduction

The COVID-19 pandemic highlighted the pressing need for adaptive forecasting models that can capture the highly nonstationary dynamics of epidemic spread. Conventional epidemiological frameworks and classical time-series models often rely on assumptions of stability in either transmission parameters or behavioral responses. However, the pandemic was characterized by abrupt shifts, which arose from the emergence of new viral variants, sudden changes in government policies, and rapid transformations in public awareness. These changes generated discontinuities that traditional models frequently failed to represent, resulting in systematic prediction errors at precisely the moments when reliable forecasts were most needed.

To overcome these limitations, our study introduces a hybrid approach that integrates heterogeneous data streams reflecting both physical mobility and discursive behavior in society. Specifically, we combine large-scale smartphone GPS mobility data with semantic signals extracted from blogs across Japan. GPS-based measures allow us to quantify collective movement patterns across categories such as home, work, move, and stay, thereby providing a proxy for contact structures that underlie viral transmission. In parallel, blog text data captures the evolving semantic networks of public concern, including shifts in attention toward health-related risks, medical system capacity, and social disruptions. We aimed to construct a robust forecasting model by utilizing these two data sources together with a methodology for extracting effective variables from large-scale candidate sets.

Methods and Results

Our empirical analysis utilizes anonymized smartphone GPS data from approximately 400,000 users in Japan. These data were classified into categories reflecting behavioral contexts: home, work, move, and stay. In parallel, we constructed a large-scale textual corpus by extracting daily frequencies of COVID-19 related terms from blogs across Japan. The resulting dataset comprised more than 13,000 candidate predictors, encompassing mobility indices, semantic features, and lagged variables. Given the high dimensionality, we designed a systematic preprocessing and selection pipeline to identify robust predictors while controlling for redundancy and spurious associations.

The pipeline involved several key steps. First, we applied trend decomposition to both mobility and blog series, enabling us to separate long-term shifts from transient fluctuations. Second, we employed correlation filtering and multicollinearity reduction to eliminate highly redundant predictors. Third, we used Bayesian approaches to further suppress spurious associations, thereby improving the interpretability and stability of the selected features. The outcome of this process can be conceptualized as a lexical-mobility network, in which nodes correspond to candidate predictors and edges represent correlations or clustering structures identified during selection. Importantly, the network structure itself exhibited dynamic evolution across epidemic waves, reflecting the shifting salience of predictors.

To evaluate predictive performance, we compared two regression frameworks. The baseline model was trained on data from the first pandemic year using a fixed training window, while the adaptive model was retrained sequentially for each epidemic wave. The adaptive model achieved approximately 10% reduction in mean absolute error compared with the baseline and reached nearly 90% trend prediction accuracy (F1 > 0.85). These improvements underscore the value of dynamically updating predictor sets in response to changing contexts.

Figure 1 illustrates the results for Japan's 5th to 7th waves. The top panel compares predicted values (green) with observed indicators (orange), along with 95% prediction intervals. Vertical blue lines mark the onset of distinct waves identified by trend decomposition. The bottom panel presents the trend decomposition of predicted and observed values, where red segments denote upward trends and blue denote downward trends. Although there is a discrepancy between the observed and predicted values in the upper panel, the trend prediction in the lower panel achieves very high accuracy.

Our study provides a flexible forecasting framework that remains robust under rapidly changing conditions, offering practical implications for public health preparedness. Representing the relationships among a large number of candidate explanatory variables as a network structure is important for understanding the complex correlations and redundancies among them. In an era where vast amounts of data are available, identifying effective predictors from numerous variables is becoming increasingly important, and the methodology proposed in this study can be applied not only to pandemic forecasting but also to a wide range of other domains.

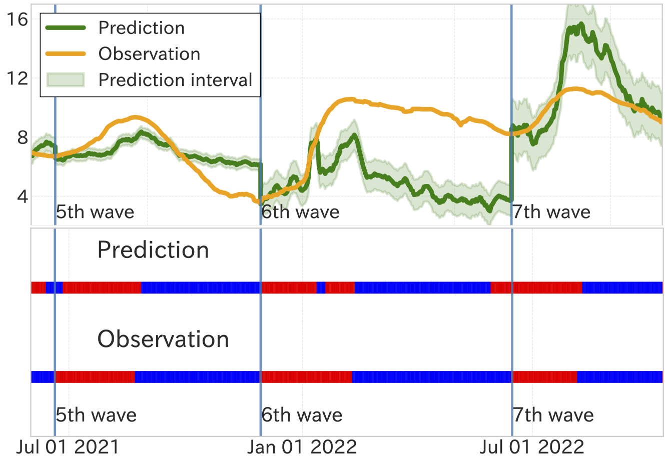

Figure 1: Results of the sequential adaptive regression model integrating epidemic forecasting with mobility and discourse data. The top panel shows observed values (orange), model ’ s predicted values (green), and prediction intervals (green bands) for Japan ’ s 5th-7th COVID-19 waves. The blue vertical lines indicate wave boundaries identified via trend decomposition. The bottom panel illustrates the trend decomposition of predicted (top row) and observed (bottom row) values, where red bands denote upward trends and blue bands denote downward trends.

References

[1] R. Susuta, K. Yamada, H. Takayasu, and M. Takayasu. A Forecast Model for COVID-19 Spread Trends Using Blog and GPS Data from Smartphones. *Entropy*, 27(7):686, 2025. `https://doi.org/10.3390/e27070686`