Forecasting COVID-19 Hospital Census: A Multivariate Time-series Model Based on Local Infection Incidence

Hieu Nguyen, MS, Philip Turk, MS, PhD, Andrew McWilliams, MPH, MD
Center for Outcomes Research & Evaluation, Atrium Health

Introduction

Background

Our work is motivated by the need from hospital leaders to have timely and accurate forecasts to guide planning for surges in hospital census, i.e., bed capacity, due to the COVID-19 pandemic. Adequate preparation can help prevent or mitigate strains on hospital resources COVID-19 that result when hospitals exceed their historical capacity.

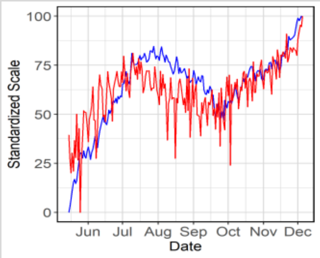
Objective

We want to explore whether the local COVID-19 infection incidence and the COVID-19 hospital census can be successfully incorporated within a multivariate time-series model to delivery satisfactory 7-day-ahead forecast performance and examine the application of this model to scenario-based long-term forecasting.

Study data

The study data are aggregated daily COVID-19 hospital census across 11 Atrium Health hospitals plus a virtual hospital in the greater Charlotte metropolitan area of North Carolina, as well as the total daily infection incidence across the same region during the May 15, 2020 – December 5, 2020 period (Figure 1). The data was applied to appropriate transformations to linearize their relationship.

Figure 1. Scaled COVID-19 hospital census and local infection incidence time-series.



Methods

Model

A Vector Error Correction model (VECM) is a vector autoregressive (VAR) model used for nonstationary multivariate time-series and accounts for stable long-run relationships, i.e., cointegration, between the time series. A time-series vector is said to be cointegrated if there is at least one linear combination of the vector that is trend-stationary.

Following [1], we first describe the VAR representation of the model, i.e., the level equation:

$$\boldsymbol{y}_t = \boldsymbol{\Pi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Pi}_p \boldsymbol{y}_{t-p} + \boldsymbol{\mu} + \boldsymbol{\Phi} \boldsymbol{D}_t + \boldsymbol{\epsilon}_t$$

for time $t = 1, \ldots, T$, where $\boldsymbol{\Pi}_i$ (for $i = 1, \ldots, p$) are k x k coefficient matrices of the lagged series at lag i, $\boldsymbol{\mu}$ is a k x 1 vector of constants, $\boldsymbol{D}_t$ is a 6 x 1 vector of weekly seasonal indicators, $\boldsymbol{\Phi}$ is a k x 6 coefficient matrix for seasonal indicators, and $\boldsymbol{\epsilon}_t$ is a k x 1 vector of random errors.

The VECM representation, i.e., difference equation, can be derived from above:

$$\Delta \boldsymbol{y}_t = \boldsymbol{\Pi} \boldsymbol{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{y}_{t-p+1} + \boldsymbol{\mu} + \boldsymbol{\Phi} \boldsymbol{D}_t + \boldsymbol{\epsilon}_t$$

Methods (cont.)

where $\Delta \boldsymbol{y}_t$ is a k x 1 vector of the differenced series $\boldsymbol{y}_t - \boldsymbol{y}_{t-1}$, $\boldsymbol{\Pi} = -(\boldsymbol{I} - \boldsymbol{\Pi}_1 - \cdots - \boldsymbol{\Pi}_p)$, and $\boldsymbol{\Gamma}_i = -(\boldsymbol{\Pi}_{i+1} + \cdots + \boldsymbol{\Pi}_p)$ (for $i = 1, \ldots, p-1$).

The model has the following assumptions:

- Assumption 1: The components of $\boldsymbol{y}_t$ are at most $I(1)$, i.e., integrated of order 1.
- Assumption 2: $0 \le r = rank(\boldsymbol{\Pi}) \le k$
- Assumption 3: $\boldsymbol{\epsilon}_t$ are identically and independently distributed $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ random vectors with covariance matrix $\boldsymbol{\Sigma}$.

For assumption 2, if $r = k$, then it can be shown that the VECM becomes a standard VAR model. If $r = 0$, then $\boldsymbol{\Pi}$ is the zero matrix and there is no cointegration relationship between the series. The VECM then becomes a VAR model for differenced time-series. If $0 < r < k$, then $\boldsymbol{\Pi}$ can be factored into $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both k x r matrices. From assumption 1, the differenced series $\Delta \boldsymbol{y}_t$, and its lags $\Delta \boldsymbol{y}_{t-1}, \ldots, \Delta \boldsymbol{y}_{t-p+1}$, are stationary. It follows that $\boldsymbol{\Pi} \boldsymbol{y}_{t-1} = \boldsymbol{\alpha}\boldsymbol{\beta}^T \boldsymbol{y}_{t-1}$, also called the error correction term, is (trend-)stationary, depending on the specification of the deterministic components. The r linearly independent columns of $\boldsymbol{\beta}$ are the cointegrating vectors and the rank r is equal to the cointegration rank of the system of time series.

Forecast performance

We used Mean Absolute Percentage Error (MAPE) to evaluate the 7-day-ahead forecasts of Census:

$$MAPE = \frac{100}{7} \sum_{i=1}^{7} \left| \frac{F_i - A_i}{A_i} \right|$$

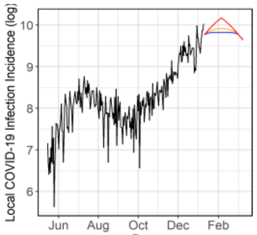
where F_i is the forecast value and A_i is the actual value.

The sampling distribution of out-of-sample MAPE is obtained by time-series cross-validation.

Long-range scenario-based forecasting

Leveraging epidemiologically informed scenarios of the future infection incidence, we attempted to use the model to create realistic projections of hospital census. On January 9, 2021, we expected the winter surge to reach peak infection prevalence around February 5, 2021 based on an extension of an epidemiological model called the Susceptible-Infected-Removed model [2]. We linearly extrapolated Incidence with positive trend up to the expected pandemic peak. The severity of a scenario was controlled by a trend-dampening parameter [3]. After the peak, the descent path was initially symmetric to its ascent and then eventually became linear (Figure 2).

Figure 2. 60-day projected local COVID-19 infection incidence of the CRI region on the log scale, as of January 9, 2021. Past values (black), worst-case scenario (red), base-case scenario (orange), best-case scenario (blue).



Results

Model estimation

- The level equation requires 7 lags (p = 7) to capture all temporal dependencies.
- Strong evidence for a cointegration relationship (*P*<.01):

$$ect_{t-1} = Census_{t-1} - 0.8013\ Incidence_{t-1} + 7.8266$$

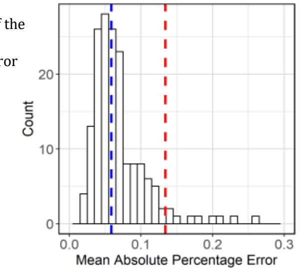
where ect_{t-1} was the (lagged) error correction term.

- Long-run effect: the error correction term had a negative and a statistically significant effect on Census change (*P*<.01).
- Short-run effects: past Incidence changes at lags 1, 2, 4, 5, and 6, as well as past Census change at lag 2, had significant effects on Census change.

Forecast performance

The typical value (median) of MAPE was 5.9% and the 95th percentile of MAPE was 13.4% (Figure 3). For the sake of comparison, the corresponding values from an Autoregressive Integrated Moving Average (ARIMA) model using the COVID-19 hospital census only were 6.6% and 14.3%.

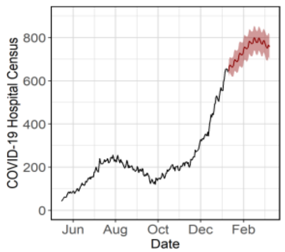
Figure 3. Distribution of the 7-day-ahead Mean Absolute Percentage Error from time-series cross-validation for the period June 16, 2020 - November 28, 2020. Median (blue), 95th percentile (red).



Long-range scenario-based forecasting

In all scenarios, due to cointegration, the hospital census followed corresponding concave trajectories with peaks occurring approximately 2-3 weeks later than Incidence depending on the scenario. In the worst-case scenario, the hospital census was projected to peak on February 16, 2021 (11 days later than Incidence) with approximately 850 patients at the 80% forecast interval upper bound (Figure 4).

Figure 4. Worst-case scenario 60-day forecasts for COVID-19 hospital census of the CRI region, as of January 9, 2021. Past values (black), forecasts (red line), 80% forecast intervals (red band).



Discussion

- We have ascertained the long-term stable relationship between local infection incidence and COVID-19 hospital census,. Whereas, current models, e.g., the COVID-19 Hospital Impact Model for Epidemics (CHIME) [4], rely on simplified assumptions about the relationship.
- Local infection incidence shows to be an effective leading indicator for COVID-19 hospital census, through both short-run and long-run effects, and as demonstrated by very good forecast performance against the traditional ARIMA model.
- In hindsight, by evaluating different scenarios of peak resource demand against our resource capacity, we have correctly assured our leaders of our capability to handle even the worst-case scenario, alleviated uncertainty, and effectively guided long-term planning of adequate staffing, bed capacity, and equipment supplies through the pandemic.

Resources

1. Pfaff B. Analysis of integrated and cointegrated time series with R. 2nd ed. New York: Springer; 2008. Wang L, Zhou Y, He J, Zhu B, Wang F, Tang L, et al.
2. An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. Journal of Data Science. 2020;18(3): 409-432. [doi: 10.6339/JDS.202007_18(3).0003]
3. Gardner ES, Mckenzie Ed. Forecasting Trends in Time Series. Management Science. 1985;31: 1237–1246. [doi: 10.1287/mnsc.31.10.1237]
4. Weissman GE, Crane-Droesch A, Chivers C, Luong T, Hanish A, Levy MZ, et al. Locally Informed Simulation to Predict Hospital Capacity Needs During the COVID-19 Pandemic. Annals of Internal Medicine. 2020;173: 21–28. [doi: 10.7326/M20-1260] [Medline: 32259197]

Contact Info

Hieu.Nguyen@atriumhealth.org

Acknowledgements

This research protocol was submitted to the Atrium Health Institutional Review Board (IRB) prior to execution and the study was deemed exempt from IRB oversight. In compliance with HIPAA regulations, individual patient information is not disclosed, all data have been deidentified and reported as aggregates.