Supplementary Material
"Computational Modeling of Regional Dynamics of Pandemic Behavior using Psychologically Valid Agents"

Peter Pirolli[1]
Institute for Human and Machine Cognition

Choh Man Teng
Institute for Human and Machine Cognition

Christian Lebiere
Carnegie Mellon University

Konstantinos Mitsopoulos
Institute for Human and Machine Cognition

Don Morrison
Carnegie Mellon University

Mark Orr
University of Virginia

Overview

The overall process for developing the Regional PVAs (Psychologically Valid Agents) is depicted in Figure S1. Data from a variety of sources, described in detail below, were collected and merged. The merged data contained predictor variables (features) for *static* regional variables (features) and time series regional variables capturing day-by-day COVID-19 statistics and mask-wearing rates. Some of the variables were subjected to dimensionality reduction (e.g., weather data) to induce a smaller set of latent features. A combination of the original variables (or standardized original variables) and dimensionality-reduced latent variables were used in a regression analysis with regional time series mask-wearing as the dependent variable. The regression outputs were used to down select a reduced set of features to serve as input drivers for the PVAs. The PVAs made time series predictions of mask-wearing dynamics in response to regional features and the effective transmission number, Rt, of COVID-19.

Figure S2 summarizes the data used in the R-PVA (Regional PVA) models. We selected a set of *static* regional variables from several sources representing data at the state and county levels of the U.S. The sources included data aggregated by the Johns Hopkins Coronavirus Resource Center[1,2], regional Big 5 personality statistics[3], and a fused data set used in a study of predictors

[1] Address correspondence to Peter Pirolli, Institute for Human and Machine Cognition, email: ppirolli@ihmc.org

of social distancing behavior[4]. Guided by previous machine learning research on the same or similar data[5], we selected a subset of variables to include in our analysis.

In addition to the static regional variables, we selected and created a set of *time series* regional variables. For each county and/or state, we combined survey data from the CovidCast and the Covid States projects. Our analysis concerns the time period of the first three waves of Covid-19 as defined by Pew Foundation[6]. R_t data at the state and county levels were downloaded from covidestim.org, a project of the Yale School of Public Health. Mask-wearing data at the state level was downloaded from the CovidStates[7]. The time series data is for every day for the pre-vaccination rollout period of 4/24/2020 to 3/31/2021. The mask-wearing data compiled by the Covid States project was captured in survey waves deployed every three weeks.

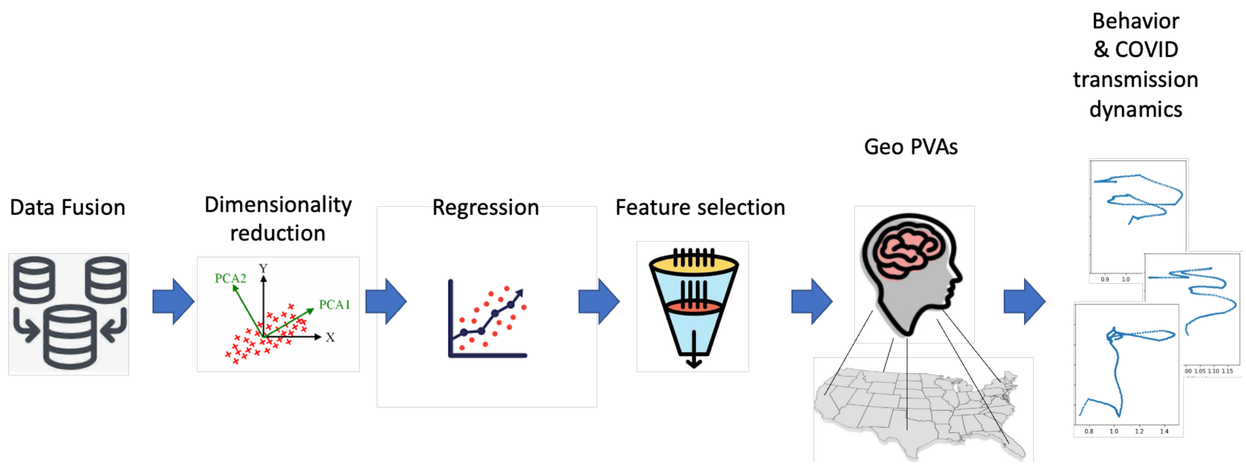

Figure S1. Overview of the data analysis and modeling pipeline.

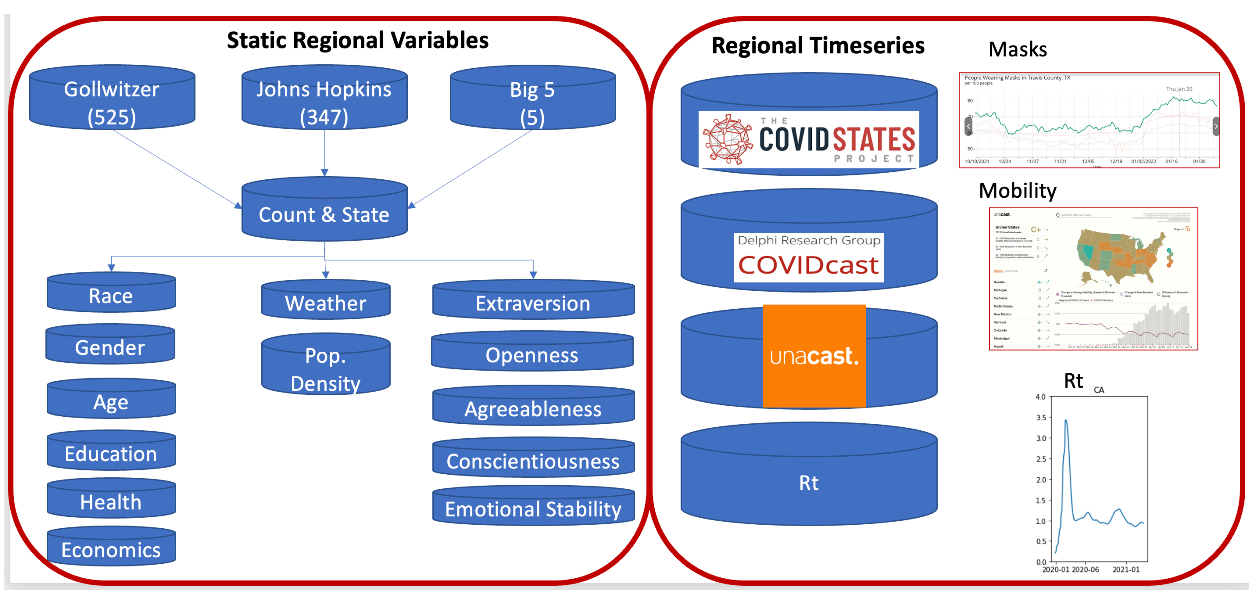

Figure S2. Overview of the data sources used in the analysis and PVA modeling. Regional static data is on the left and time series data is on the right.

Data

Details of the specific predictor variables and their sources (Figure S2) are presented in Table 1. The raw version of the variables and standardized z-score values of the variables were explored in the modeling and analysis. The selection of this initial set of variables was motivated by prior work by Nicholson et al.[5] who applied machine learning and clustering techniques to data similar to those in Table 2 along with county-level mobility data during the early phases of COVID-19.

Our dimensionality reduction of the raw weather data provided by Johns Hopkins to produce the latent principal components in Table 1 also followed the Nicholson et al. approach. Using PCA (Principal Components Analysis), the weather data was first mean-centered and scaled with respect to feature standard deviation. Next, the data was projected from 48 dimensions to 2 principal components while retaining approximately 78% of the original variation. The first principal component (PC1) explains 65% of the variance and is dominated by the monthly temperature related variables. The second principal component (PC2) explains 13% of the variance and is dominated by the monthly precipitation related variables during the fall and winter (Oct–Apr).

Table S1. Static regional variables

Type	Variable	Description	Source
Weather	PC1_weather	Derived weather temperature factor based on PCA	Johns Hopkins
	PC2_weather	Derived weather fall-winter precipitation factor based on PCA	Johns Hopkins
Race/ethnicity/gender	NHWA	Percentage of Non-Hispanic, White, living alone	Johns Hopkins
	NHBA	Percentage of Non-Hispanic, Black, living alone	Johns Hopkins
	NHIA	Percentage of Non-Hispanic, American Indian, living alone	Johns Hopkins
	TOM	Percentage of Two or more races	Johns Hopkins
	Gender ratio	Total male/total female population	Johns Hopkins
Age	Percentage age 0-17 yr		Johns Hopkins
	Percentage age 65+ yr		Johns Hopkins
Education	Percentage adults with less than a high school diploma		Johns Hopkins
	Percentage adults with a bachelor's degree or higher		Johns Hopkins
Economics	Median household income 2018		Johns Hopkins Johns Hopkins
	Unemployment rate 2018		Johns Hopkins
Media Diet	Fox News Lean	Proportion of people reporting watching Fox News minus proportion of people reporting	Gollwitzer

		watching CNN and MSNBC, 2019	
Politics	GOP Advantage 2016	Percent Republican vote minus percent Democrat vote, 2016	Gollwitzer
	PctTrump State 2016	Percent Trump vote, 2016	Gollwitzer
	Trump approval 2020	Trump net approval (i.e., approval minus disapproval), -1 to 1 scale, February 2020	Gollwitzer
Population	Population density per square mile		Gollwitzer
Personality (Big 5)	Extraversion	Talkative, energetic, assertive, and outgoing.	Ebert
	Openness	Willing to listen to multiple viewpoints or try new things. Those who are lower in openness tend to be averse to change and skeptical of new ideas.	Ebert
	Agreeableness	Willing to listen to multiple viewpoints or try new things. Those who are lower in openness tend to be averse to change and skeptical of new ideas.	Ebert
	Conscientiousness	Organized, reliable, and ambitious	Ebert
	Emotional Stability	ability to cope with stress, resist our impulses, and adapt to change. People who score high on emotional stability tend to be calm, composed, and stress-resistant.	Ebert

We used the mask-wearing survey data provided by the CovidStates project. This is because the surveys began in April 2020. The alternative was to use data from CovidCast, which did not begin until September 2020. The actual survey question used in the CovidCast survey was:

- In the last week, how closely did you follow the health recommendations listed below? (Check all that apply)
    - Wearing a face mask when outside of your home
    - Frequently washing hands
    - Avoiding contact with other people
    - Avoiding public or crowded places

Whereas in the CovidStates survey the question was:

- In the past 5 days, how often did you wear a mask when in public?

In both surveys, a data reweighting was applied to the samples to produce a probability weighting representative of the U.S. demographics.

The disadvantage of using the CovidStates mask-wearing data is that the surveys were conducted in 3-week rolling waves, whereas the CovidCast data were collected every day. For the purposes of our analysis and modeling we performed an interpolation to produce state-by-state, day-by-data mask-wearing estimates from the CovidStates data. For each state, we ordered the survey data by the date that occurred mid-survey wave. We iterated through each pair of consecutive mask-wearing data in that ordered list. For each pair of observations, we iterated through each date between the lower-dated and upper-dated observations and performed a linear interpolation weighted by the ratio between elapsed days from the lower-dated value and days remaining to the upper-dated value.

Figure S3 shows the scatterplot between the CovidStates interpolated data and CovidCast observed data, where it was available. Figure S3 suggests that the CovidStates interpolated data are approximately 15 points lower than the CovidCast data. A linear regression suggests a high coefficient of determination or $R^2 = 0.87$.

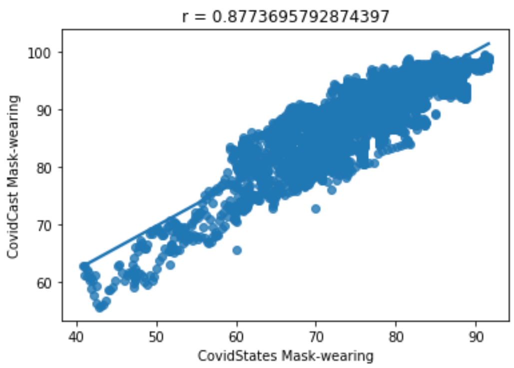

Figure S3. Scatterplot of the CovidCast mask-wearing data against the interpolated CovidStates data.

Feature Selection using Stepwise Regression

As discussed in the main article, we performed stepwise regression to select features. We used one day of data at the mid-point of the time series (2020/10/11) for the evaluation. Starting with an empty set of variables, at each step a regression model was constructed with the current set of selected variables plus one of the unselected variables to predict mask-wearing. The unselected variable that resulted in the best average R^2 in a 10-fold cross validation was added to the selected set. Figure S4 shows the coefficients of determination obtained for regression models as additional features were added to the regression. The successive scores were inspected and a cutoff of 5 variables was deemed the optimal balance between the number of variables and the resulting score. The 5 variables are 'PctTrump State 2016', 'PC1_weather', 'Fox News Lean', 'Percentage adults with a bachelor's degree or higher' and 'Percentage age 65+ yr'.

Figure S4 shows the R^2 scores of the regression models built at each stage of the selection process.

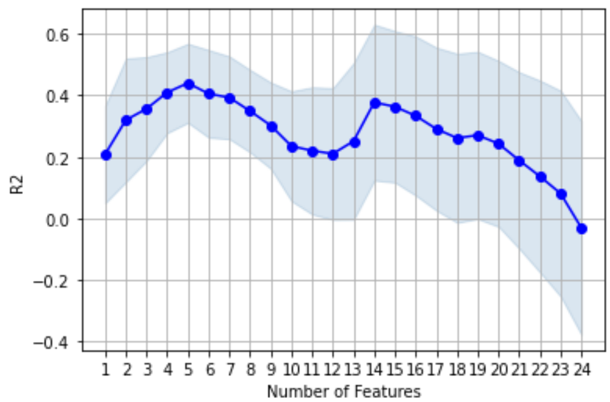

Figure S4. R^2 scores of regression models built with successively larger sets of features

Figure S5 shows RMSE scores obtained for different combinations of effort and boost factor. As noted in the main article, we performed a grid search over the parameter space of effort and boost factor, with the value of actual difficulty fixed at 2.0. We used only the data for two states, California and Wyoming, in the date range 2020/04/24-2020/06/30, corresponding to the first wave of COVID-19. *n*-fold rolling origin cross-validation was used for scoring the models. The best RMSE was 0.0369, for boost factor = 0.02 and effort = 1.0

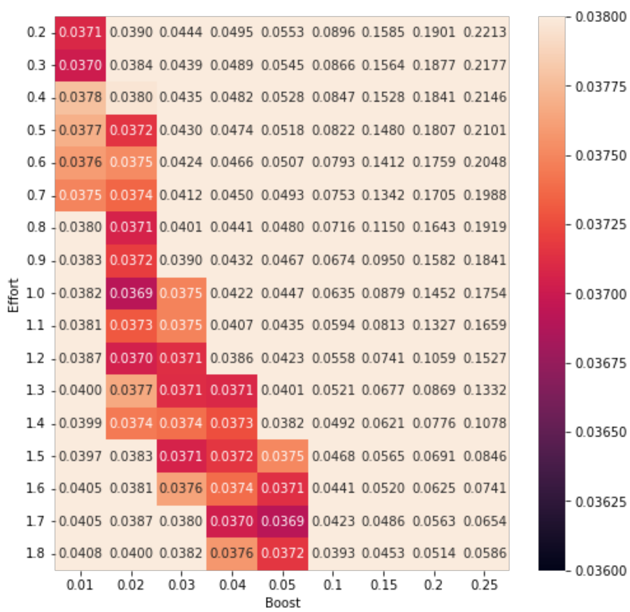

Figure S5. Grid search over combinations of boost factor and effort, scored using RMSE. The best RMSE is 0.0369, for boost factor = 0.02 and effort = 1.0.

We compared the performance of the four R-PVA models in Figure S6. In general, the predictions were more accurate as more features were included and also when the self-efficacy process was modelled.

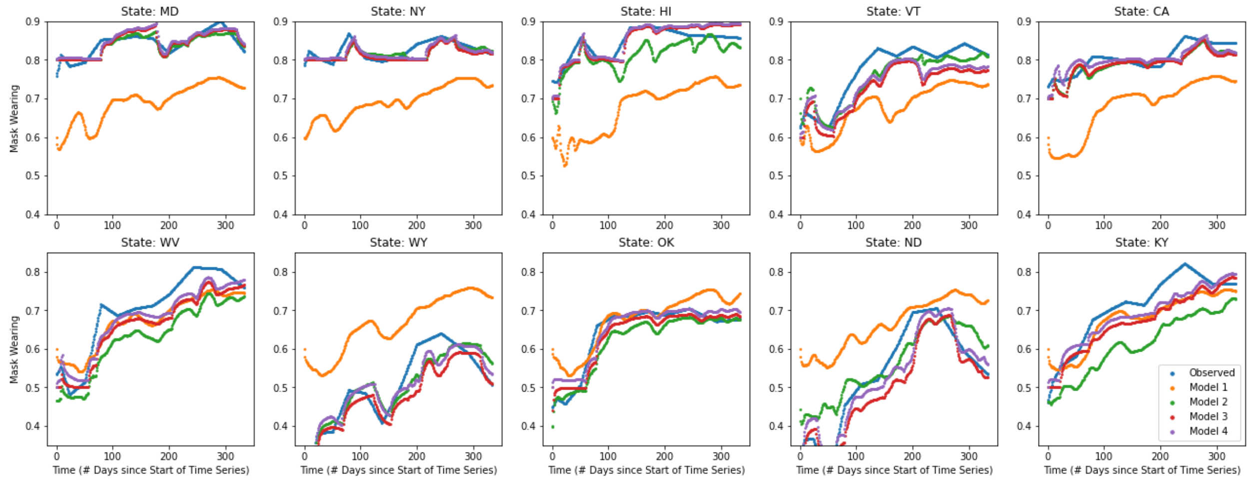

Figure S6. Observed proportion of mask wearing and proportion of predicted by the four R-PVA models.

We plotted each of the Big Five features against mask wearing for the states with the lowest and highest average mask wearing percentages (denoted by lowMask and highMask), as well as for the lowTrump and highTrump states. Figure S7 shows the values with the regression line.

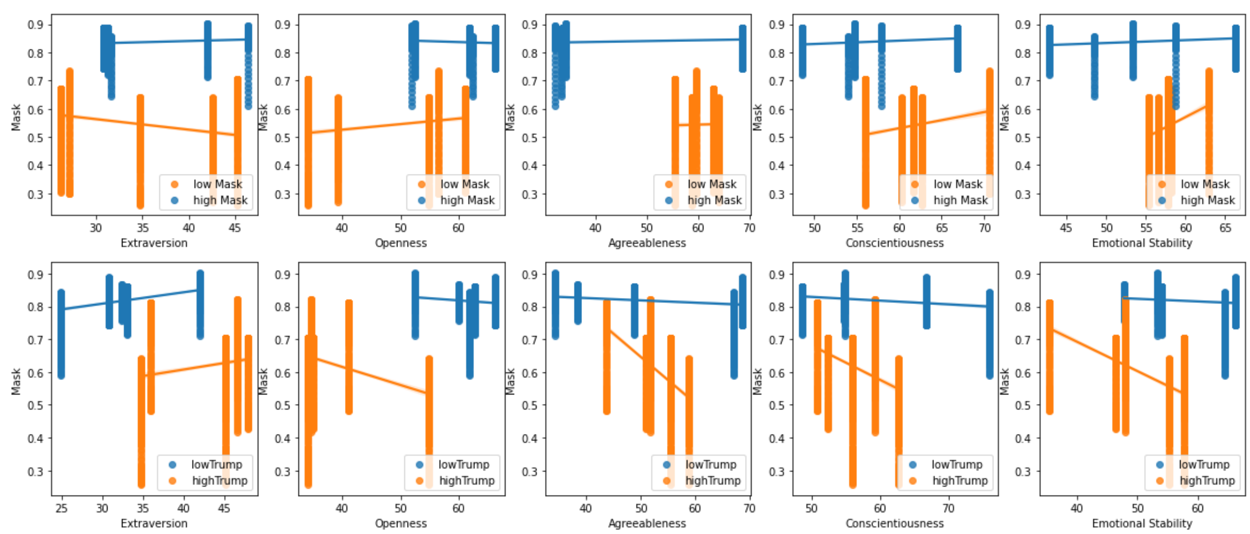

Figure S7. Linear regression of Big 5 features against mask wearing, for the states with low and high mask average wearing percentages, and for lowTrump and highTrump states.

References

1 Center, J. H. C. R. (ed Johns Hopkins University and Medicine) (n.d.).

2 Killeen, B. D. *et al.* A County-level Dataset for Informing the United States' Response to COVID-19. arXiv:2004.00756 (2020). <https://ui.adsabs.harvard.edu/abs/2020arXiv200400756K>.

3 Ebert, T. *et al.* Are Regional Differences in Psychological Characteristics and Their Correlates Robust? Applying Spatial-Analysis Techniques to Examine Regional Variation in Personality. *Perspectives on Psychological Science* **17**, 407-441 (2022). https://doi.org/10.1177/1745691621998326

4 Gollwitzer, A. *et al.* Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature Human Behaviour* **4**, 1186-1197 (2020). https://doi.org/10.1038/s41562-020-00977-7

5 Nicholson, C., Beattie, L., Beattie, M., Razzaghi, T. & Chen, S. A machine learning and clustering-based approach for county-level COVID-19 analysis. *PLoS One* **17**, e0267558 (2022). https://doi.org/10.1371/journal.pone.0267558

6 Jones, B. *The changing political geography of COVID-19 over the last two years*, <https://www.pewresearch.org/politics/2022/03/03/the-changing-political-geography-of-covid-19-over-the-last-two-years/> (2022).

7 Lazer, D. *et al.* The COVID States Project #26: Trajectory of COVID-19-related behaviors. (2021).