Towards a data-driven characterization of behavioral changes induced by the seasonal flu

Nicolò Gozzi\textsuperscript{1, }*, Daniela Perrotta\textsuperscript{2}, Daniela Paolotti\textsuperscript{3}, Nicola Perra\textsuperscript{1, 3}

\textsuperscript{1} Networks and Urban Systems Centre, University of Greenwich, London, UK
\textsuperscript{2} Max Planck Institute for Demographic Research, Rostock, Germany
\textsuperscript{3} ISI Foundation, Turin, Italy

* N.Gozzi@gre.ac.uk

S1 Appendix

Descriptive analysis of data

We report here a few qualitative observations on the data. First, responses are divided in the three behavioral categories as follows: 26% report no changes, ii) 36% only moderate changes, iii) 38% significant variations. Thus the majority of them describe either moderate or severe changes in behaviors.

Second, only 20% of the participants that reported variations in their behaviors in a particular season, experienced the flu in that season before or while compiling the survey. We identify the cases of flu using the definition of the European Centre for Disease Prevention and Control (ECDC)\textsuperscript{[1]}. According to this definition, seasonal flu is identified by the following criteria: sudden onset of symptoms, at least one systemic symptom (fever or feverishness, malaise, headache, myalgia), and at least one of respiratory symptom (cough, sore throat, shortness of breath).

Third, during 2017—18 season, 27% of the answers provided before the peak are related to the category of no change, 46% to moderate change and 27% to social distancing. After the peak, instead, the number of surveys related to no change is 23%, to moderate change is 34%, while the percentage of surveys related to social distancing significantly grows up to 43%. Very similarly, during 2018—19 season before the peak we have 29% of surveys that report any change in behavior, 39% that report only moderate change to behavior, and 32% that report the adoption of social distancing measures. The division after the peak, instead, is the following: 26% of surveys related to no change, 29% to moderate change and 45% to social distancing (Fig A). Interestingly, in both seasons we note a significant increase in the number of surveys reporting social distancing measures after the peak. This observation highlights how risk perception is linked to the progression of the disease and not constant throughout. Indeed, proactive behavioral measures adopted by people seem to strengthen when external conditions get worse.

Fourth, the comparison of our sample with the statistics retrieved from the online archive of the Italian National Institute of Statistics (ISTAT)\textsuperscript{[2]} in Fig B shows that the geographic, gender and age distributions of participants are almost in line with the Italian population as a whole. In particular, we note a slight over representation of i) Northern Italy (+13% north west, +7% north east), ii) of male (+11%), and iii) of age group 15-64 (+12%). In Fig C we compare also our sample to the general population in terms of education level. In this case we observe a significant over-representation of individuals with a higher level of education (+33%). We know from previous research
Flu incidence in the two seasons (above) and behavioral class distribution before/after the peak (below). The two seasons had a similar impact in terms of epidemic incidence, while the peaks are slightly shifted (2nd week of 2018 for the 2017/18 flu season, and 5th week of 2019 for the 2018/19).

Comparison of our sample with the general Italian population in terms of age, gender and home area.

that this can be an issue associated with internet-based studies [3–5]. The representativeness of Influweb - the platform we used to collect data - has been addressed in more detail in [6]. Unfortunately, we do not have information regarding the level income of participants. However, previous research on participants of Flu Near You and of other Influenzanet campaigns showed that their socioeconomic status tends to be higher than the general populations in their respective countries [7–9].

Finally, in table A we list the behaviors changed by individuals in the class of social distancing and in the class of moderate change. As an example, we classify in the class...
Fig C. Comparison of our sample with the general Italian population in terms of education level.

of social distancing an individual who responded both “Reduced the amount you go to school/work” and “Reduced the amount you touch your eyes, nose, or mouth”, while we classify in the class of moderate change an individual who responded both “Cleaned or disinfected things you might touch more often” and “Usually carried tissues with you when out more often”.

Table A. Behaviors changed in social distancing and moderate change classes.

<table>
<thead>
<tr>
<th>social distancing</th>
<th>moderate change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced the amount you go to school/work</td>
<td>Cleaned or disinfected things you might touch more often</td>
</tr>
<tr>
<td>Cancelled or postponed social events such as meeting friends</td>
<td>Carried sanitising hand gel when going out more often</td>
</tr>
<tr>
<td>Reduced or changed the way you use public transportation</td>
<td>Used sanitising hand gel to clean your hands more often</td>
</tr>
<tr>
<td>Reduced the amount you go into shops</td>
<td>Reduced the amount you touch your eyes, nose, or mouth</td>
</tr>
<tr>
<td>Kept one or more of your children out of school or nursery</td>
<td>Followed a healthy diet or took vitamin supplements more often</td>
</tr>
<tr>
<td>Kept away from crowded places</td>
<td>Usually carried tissues with you when out more often</td>
</tr>
<tr>
<td>Tried to avoid people who have flu</td>
<td>Usually used tissues when sneezing or coughing more often</td>
</tr>
</tbody>
</table>

Classification algorithms

Decision tree ensembles are a powerful tool for classification tasks [10]. They consist in a set of classification trees. The underlying idea is that summing together the predictions of multiple “weak” learners, one can achieve more robust predictions than with a single “strong” learner. Mathematically, the prediction $\hat{y}_i$ of this model for the $i$-th point of the dataset can be written as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$  (1)

Where $K$ is the number of trees (i.e. the size of the ensembles) and $f_k(x_i)$ is the prediction of the $k$-th tree for the $i$-th point. This general model is implemented by a
great variety of algorithms, such as Gradient Boosted Trees (GBT) \cite{11} and Random Forest (RF) \cite{12}. The differences arise from how trees are built and added. GBT exploits a specific training strategy called additive training, in which at each training step is added to the ensemble the tree that optimizes the objective function. Recently, GBT has gained a great popularity for its speed and performance, and has become the algorithm of choice for many machine learning applications.

**Support Vector Machines** (SVM) \cite{13} extends the concept of Linear Classification by trying to find a hyperplane in an N-dimensional space (where N is the number of features) that distinctly classifies the data points and that has the maximum margin. This approach is more stable to noise and to the classification of future points. However, one of the main advantages of SVM is the so-called “kernel trick”. In fact, using this method SVM can easily handle also seriously non-separable data. Through the kernel trick data can be easily projected in a vector space where they are - eventually - linearly separable. The only request for this new space is that the dot product (called kernel) must be defined. Commonly used kernels are the polynomial kernel or the radial basis function kernel (RBF). This approach makes SVM a very powerful tool that can be exploited to classify a broad range of datasets.

**Logistic Regression** (LG) \cite{14} is a very common tool used to predict probabilities. Given a sample input vector \( x \in \mathbb{R}^n \) and the vector of model’s weights \( \theta \in \mathbb{R}^n \), LG smoothly projects the signal \( s = \theta^T x \) in the probability range \([0, 1]\) using - for example - a logistic function defined as:

\[
\gamma(s) = \frac{e^s}{1 + e^s}
\]

The output can then be interpreted as a probability for a binary event. This framework can be easily extended to the cases with more than two possible outcome.

**Dummy Predictors** (RND) are generally used as null benchmarks to assess whether the models learnt something from data. There are many trivial prediction strategies, for example constant or equally probable predictions (“coin toss”). In this work we consider a random predictor that generates random predictions by respecting the training set’s class distribution.

For all the models we use open-source implementation. In particular, we use XGBoost, an open-source software library which provides a “scalable, portable and distributed implementation” of GBT \cite{15,16}. For other models we use the implementation provided by the open-source library scikit-learn \cite{17} and we train them fine-tuning standard parameters. The source code can be find at \cite{18}.

**Sensitivity analysis**

We test here the stability of classification results of GBT. To this end, we train the model on various subsets of the whole dataset. First, we classify separately the surveys submitted during the 2017−18 (\( N = 331 \) surveys) and the 2018−19 (\( N = 268 \) surveys) flu season. We obtain a precision of 0.572 for the first and of 0.474 for the second season. These results are still significantly higher than the random prediction and are comparable to those obtained in the general case with the two seasons aggregated (precision = 0.546). Second, we repeat the classification excluding from the set of features those extracted from behavioral surveys. The goal of this analysis is to explore the possibility of inferring attitudes towards behavioral change on a much bigger scale. In fact, behavioral questionnaires represent an additional burden for the participants and mostly important Influweb is one of 9 platforms in Europe. As reported in table \ref{table:B}, disregarding features deriving from behavioral surveys leads to a precision of 0.435. Despite an expected decrease in model’s performance, results are still better than the random guess. These observations are encouraging and shed light on the possibility of
extending of our approach. However, larger samples as well as data from other countries are needed to reinforce and scale our findings. Third, we consider a much simpler definition of the most important feature used in classification: disease score. In particular we consider the average number of symptoms reported by different individuals in symptoms questionnaires as a measure of severity of past illnesses. In other words, we disregard the exponential temporal weights of the original definition. Looking at results in table [B] we note that the adoption of this simpler definition of disease score does not significantly affect results, even if the classification performance across the four metrics is slightly lower. Fourth, we repeat the analysis considering only one survey for each user (we consider the most recent ones). In table [B] we can see that the classification performance is still higher than the random benchmark. Finally, we consider 2 behavioral classes instead of 3. The problem then become a binary classification between the class of no change and that of any change. We report a precision of 0.743 with respect to 0.618 of the random prediction. Notably, the five most important features in the unique and 2 classes cases are still those emerged in the complete 3-class problem (perceived susceptibility, disease score, info seeking, perceived severity and age).

Table B. GBT sensitivity analysis.

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>bal. accuracy</th>
<th>recall</th>
<th>f1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RND</td>
<td>0.343</td>
<td>0.335</td>
<td>0.334</td>
<td>0.335</td>
</tr>
<tr>
<td>2017/18</td>
<td>0.572</td>
<td>0.556</td>
<td>0.550</td>
<td>0.553</td>
</tr>
<tr>
<td>2018/19</td>
<td>0.474</td>
<td>0.455</td>
<td>0.444</td>
<td>0.441</td>
</tr>
<tr>
<td>no behavioral survey</td>
<td>0.435</td>
<td>0.432</td>
<td>0.439</td>
<td>0.435</td>
</tr>
<tr>
<td>simplified disease score</td>
<td>0.530</td>
<td>0.533</td>
<td>0.528</td>
<td>0.524</td>
</tr>
<tr>
<td>unique</td>
<td>0.478</td>
<td>0.461</td>
<td>0.473</td>
<td>0.472</td>
</tr>
<tr>
<td>2 classes</td>
<td>0.743(0.618)</td>
<td>0.646(0.504)</td>
<td>0.761(0.503)</td>
<td>0.748(0.532)</td>
</tr>
</tbody>
</table>

SHAP, an example

In SHAP, the classic definition of the Shapley value is adapted to the specific problem of assigning to each feature its importance in the classification task. The result is called SHAP value and, for feature \( i \), is defined as:

\[
\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (M - |S| - 1)!}{M!} \left[ p(S \cup \{i\}) - p(S) \right] \]  

(3)

Where, \( S \) are all the possible subsets of features without \( i \); \( |S| \) is the number of features in \( S \); \( M \) is the total number of features; \( p(S \cup \{i\}) \) is the outcome of prediction considering both \( i \) and features in \( S \), and \( p(S) \) is the outcome considering only features in \( S \). To better understand the meaning of \( \phi_i \), we divide its expression in blocks. In expression (3) part (1) is an interaction term that accounts for all possible sequences in which features are added: in fact, the order in which features are taken into account is meaningful to how their importance is assigned. Given this, we move to part (2). Intuitively, if feature \( i \) has a negligible impact on model’s decisions, then considering it or not should not affect the prediction. Looking at (3) this is equivalent to saying:

\[
p(S \cup \{i\}) \simeq p(S) \rightarrow \phi_i \simeq 0
\]  

(4)

On the other hand, if feature \( i \) is decisive for the classification we expect \( p(S \cup \{i\}) \neq p(S) \) and \( \phi_i \neq 0 \). As a guiding criterion, if feature \( i \) is important for the
classification of an individual then its SHAP value will be significantly different from zero. More in detail, if it is negative, the feature under consideration lowers the likelihood of belonging to that particular class, if it is positive it raises it.

In this appendix we present also an example on how SHAP works and how it helps in understanding model’s decisions. In Fig D we try to explain the class prediction for a specific individual made by GBT. The chart is composed of three figures and each one refers to a specific class: D1 is related to the no change class, D2 to moderate change, and D3 to social distancing. The individual under examination belongs to the class of social distancing and is classified correctly. This can be figured out examining the output of the model for each class: $-0.91$ for no change class, $-0.87$ for moderate change, and $2.50$ for social distancing. Intuitively, these numbers indicate that the most unlikely class for this individual is no change. A little bit more likely is moderate change, but the class of social distancing is far more probable. However, SHAP provides us much more information. In fact, we can evaluate the role played by each feature in the classification: the blue ones in figure have a negative effect on the likelihood of belonging to that particular class, while the red ones have a positive effect. Furthermore, the length of the arrow (red or blue) related to each feature represents its SHAP value. Then, for example, we can conclude that for this individual, info seeking and perceived susceptibility both have an important negative impact on the likelihood of being in the no change class. Conversely, they have a great positive impact on the likelihood of being in the correct class, social distancing. Finally, we can give an interpretation of the prediction: since the individual feels vulnerable to the disease (perceived susceptibility = 4) and searches regularly for information regarding the disease (seek = 1), then she is with great confidence a member of the class of social distancing.

Fig D. SHAP example: 1) is related to no change, 2) to moderate change, and 3) to social distancing class.
References


2. Istat Online Archive; https://www.istat.it/it/archivio/demografia


16. XGBoost repository: https://github.com/tqchen/xgboost


18. Source code: https://github.com/ngozzi/behavioralchange