Feature Selection for Real-time Estimates of Influenza-like Illness

Jun Li
Department of Physics
Carnegie Mellon University
junl2@andrew.cmu.edu

1 Introduction

Influenza epidemics causes significant health issues and economic burden to the society. It is estimated that the annual epidemics result in 3 to 5 million cases of severe illness, and about 250,000 to 500,000 deaths worldwide (WHO, 2016). Economically, the influenza epidemics contribute a $87.1 billion burden on the United States (US) alone each year (Molinari et al., 2007). One way to reduce the losses is to provide accurate, reliable forecasts of the influenza epidemics. Short-term forecasts within a season can help policy makers tailor the vaccine campaign, and guide individuals and organizations to adjust their activity plans to curb the influenza transmission. Long-term predictions can provide valuable information on selecting vaccines for future seasons (Brooks, Farrow, Hyun, Tibshirani, & Rosenfeld, 2015). With well-documented surveillance data for influenza-like illness, and various other digital surveillances such as search engine and social network data available nowadays, these goals seem feasible.

Reliable forecasting of influenza strongly depends on timely, accurate estimates of influenza prevalence at present and in the past. In the US, the often used gold standard of influenza surveillance is population-weighted percent influenza-like illness (wILI), derived from the US Outpatient Influenza-like Illness Surveillance Network (ILINet) (CDC, 2016). Influenza-like illness (ILI) is defined as fever and a cough and/or a sore throat without a known cause other than influenza. Each week, the health providers who volunteered to participate in the ILINet program submit percentages of ILI in their cases. The US Centers for Disease Control and Prevention (CDC) process the data and publish wILI weekly for each of the 10 HHS regions as well as on the national level. While the wILI data released by CDC offers valuable information on influenza prevalence in the US, it has two notable deficiencies (Farrow & Rosenfeld, 2017). First, there is an intrinsic, one-week delay of the data: the initial report of wILI for this week is not available until next Friday. Second, after the inital report, the wILI is subject to significant revisions in the subsequent weeks, even months as updates continue to arrive — a phenomenon known as back filling. It is because of these aspects that (Farrow, 2016; Farrow & Rosenfeld, 2017) developed their "Nowcast" methodology, which attempts to overcome the wILI's shortcomings and obtain accurate, real-time estimates of ILI by combining diverse sources of information. They have shown that the combined estimate by their Nowcast is significantly more accurate than estimates based on any one of the sources alone.

While the Nowcast provides a good way to combine different data sources for influenza nowcasting, it does not tell or specify which source should be included and which should not. In this work, we focus on providing some insights on how to select data sources for influenza nowcasting. More accurately, we study how to select from new data sources to be added to the Nowcast. The data we are going to use is sales data of grocery products in the US. We treat sales of each product as a feature. In addition, data that are being used by the existing Nowcast (Farrow, 2016; Farrow & Rosenfeld, 2017) is handled appropriately. We review several commonly used feature discovery and selection techniques in the literature. The data that we use is representative in terms of different degrees of sparsity, granularity (in both time and space), and correlation (between features).

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Problem Statement

In this work, we aim at improving the accuracy of real-time estimates of influenza-like illness with additional data sources as sensors. Specifically, we study how to select a subset of sales data of grocery products such that when combined with the data sources that are currently being used by the Nowcast will achieve the maximum improvement. We look at and experiment with various feature discovery and selection techniques, including correlation analysis, orthogonal matching pursuit (OMP), least absolute shrinkage and selection operator (LASSO), heuristic selection based on weighted linear regression (HSWLR), and Nowcast-based sequential feature selection (NSFS). A sensor is made out of the subset of features selected by each technique and fused with other data sources in the current Nowcast system (Farrow & Rosenfeld, 2017). Evaluation based on realistic nowcasting is performed for each of the feature selection methods.

3 Background and Related Work

As Gold Standard for national and regional weighted ILI, we use the CDC-reported "final" values of wILI, typically defined as the values available at week 30 of a year (late July — well after the end of the flu season). This is because updates continue to arrive many weeks after the initial report. Using the Gold Standard, we can evaluate accuracy of CDC-wILI as a function of time since it is initially published.

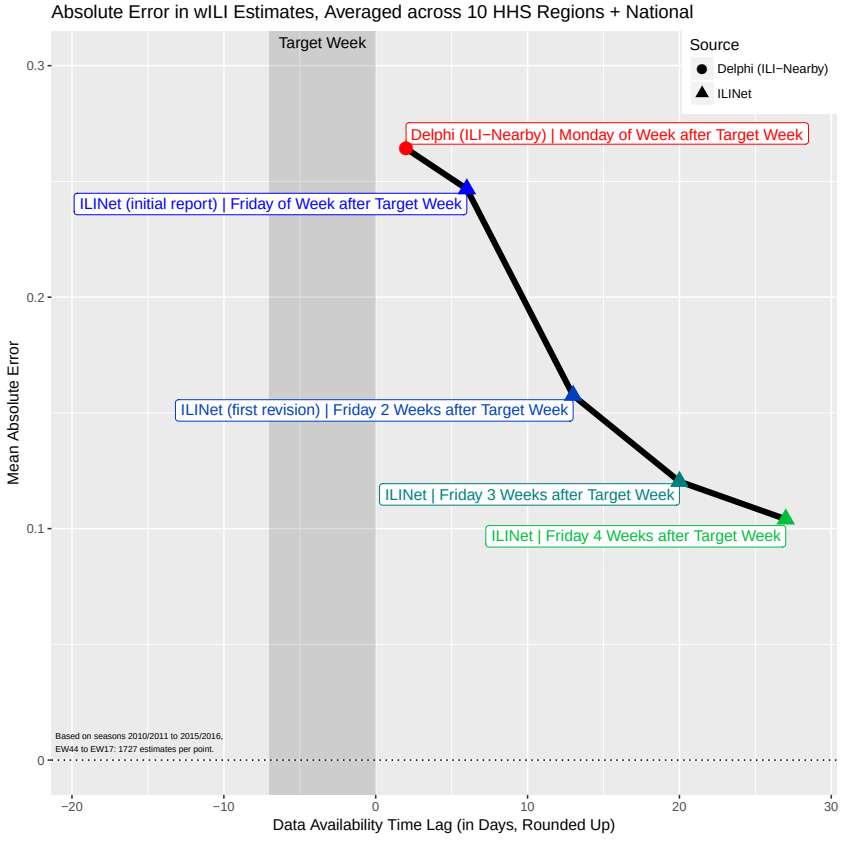


Figure 1: Mean absolute error of estimates of influenza-like illness. Evaluation is based on seasons from 2010/2011 to 2015/2016. (credits: Logan Brooks)

Figure 1 shows the mean absolute error (MAE) of the first several versions of wILI. We see that errors of the wILI of the first few versions are quite significant, and as updates continue to be made, the errors keep going down. We also note that the change between the initial report and the first revision is most dramatic. For illustration and comparison, we have also included the MAE of the Nowcast currently used (Farrow & Rosenfeld, 2017). The Nowcast generates estimates of wILI four

days ahead of initial CDC-wILI report each week, but only with a slightly worse accuracy. It should be emphasized that the nowcasts are all performed "out of sample", without knowledge of anything that was not known at nowcasting time.

For epidemiological forecasting and nowcasting, a wide variety of models have been explored and employed. Generally they can be put into three categories: compartmental models, agent-based models and parametric statistical models (Brooks et al., 2015). The compartmental models typically involve assuming some mechanistic model governing the transition dynamics of hidden variables (Brauer, Castillo-Chavez, & Castillo-Chavez, 2001; Newman, 2002). For example, the susceptible-infected-recovered-susceptible (SIRS) model approximates the dynamics of the population susceptible to influenza, the population infected with influenza, and the population recovered from influenza. Common assumptions made for the SIRS model include fully mixed population and identical transmission behavior for different strains of influenza (Shaman & Karspeck, 2012). The agent-based models build complex schemes of interaction and disease behavior in synthetic populations, and are commonly applied to the special case of a single, novel influenza strain (Ferguson et al., 2006; Colizza, Barrat, Barthelemy, Valleron, & Vespignani, 2007). The parametric statistical models are mostly inspired by the time series analysis. For example, trend of influenza prevalence can be captured by applying a linear autoregressive model to the influenza prevalence in the recent past. More complex models can be considered, such as Box-Jenkins analysis and seasonal autoregressive integrated moving-average models (Shumway & Stoffer, 2013). More recently, (Brooks et al., 2015) takes a nonparametric approach, Empirical Bayes, that generates possibilities for the current season's epidemic curve using modified versions of past seasons' curves. This method has the advantage that it outputs a distribution over epidemic curves and can thus be used to provide information on point predictions of the influenza, confidence interval and log likelihood score of the predictions. In contrast to the Empirical Bayes that aim for long term (e.g. an entire season) forecasting of influenza, the Nowcast by (Farrow & Rosenfeld, 2017) is focused on making short term (including at present) estimates of influenza as accurate as possible.

Feature selection has gained tremendous popularity over the past decade, partly to due emergence of "big data" — availability of huge amounts of data to researchers and industrial users. The purpose of selecting a subset of features instead of throwing them all in some model is two fold: to improve prediction accuracy and to obtain better interpretation (Hastie, Tibshirani, & Friedman, 2009). There are many different ways to select the features. Probably the simplest way is the one based on the correlation criteria. Intuitively, we want to select features that show as high a correlation with the response variable as possible. More sophisticated, the best subset selection finds a subset of size $k \in \{0, 1, 2, \ldots, p\}$ from p features that produces the least squared error, $\|y - X\beta\|_2^2$ (Beale, Kendall, & Mann, 1967; Hocking & Leslie, 1967). Forward stepwise selection is an iterative greedy algorithm that at each iteration it adds in a variable (feature) that best improves the fit (Efroymson, 1966; Draper & Smith, 1966). When the fit is defined as the correlation between a variable in the not-yet-selected subset and the residual after projecting the response vector y onto the selected subset of features, the algorithm is also known as orthogonal matching pursuit in the field of signal processing (Cai & Wang, 2011). The LASSO is another popular strategy for sparse feature selection (Tibshirani, 1996; Chen, Donoho, & Saunders, 2001). It is the ℓ_1 norm regularized regression that can be viewed as a convex relaxation of the non-convex best subset selection problem.

4 Data

The data we are mainly going to use is retail sales data of grocery products provided by a private company. The data contains sales data of grocery products from all major retailers in the US in 45 states. The data consists of three datasets, of different spatial and temporal granularities. We list them below along with a brief description.

- *Full-categories*

  A list of $\sim$1350 product categories, varying from ORANGE JUICE to CANNED SOUP to PERSONAL THERMOMETER. The categories are in a multi-level hierarchy. Each category comes with monthly sales volume (in Units of US dollars) for the past year on the national level (13 data points for each category).

  After receiving this dataset, we performed some simple correlation analysis. We selected two subsets (of different sizes with one being more aggressive than the other) of categories

based on the correlation with wILI (e.g. $|\rho| > 0.5$) and domain knowledge (e.g. excluding very unlikely categories such as FZ PIZZA CRUSTS/DOUGH). We then contacted the company and asked for more data for these two subsets of categories, in terms of either longer time span or finer geographical resolutions. Consequently we ended up with the following two datasets.

- *Flu-117*
  Weekly, national sales data (units and volume) from 2012 to 2017, for $\sim$120 product categories. This is for the bigger subset of categories we selected from *Full-categories*.
- *Priority-List*
  Weekly sales data (units and volume) from 2012 to 2017 on the state level for 45 states for 21 categories. This is for the smaller subset of categories we selected from *Full-categories*.

Since the goal is to improve accuracy of the existing Nowcast and the major contribution to the Nowcast is SAR3 (seasonal autoregression) data, we also uses the SAR3 data in the rest of this work. The SAR3 is a regression model with three autoregressive covariates of wILI in the recent past, two harmonic covariates for seasonal effect of the influenza and four indicator covariates accounting for holiday effects (Farrow, 2016; Farrow & Rosenfeld, 2017). Historical SAR3 data are all saved and can be retrieved from the server.

5 Method

There is a wide variety of feature discovery and selection methods in the field of statistics and machine learning. Here we discuss some of them that are commonly used and have been applied to our data.

5.1 Correlation Analysis

Let us denote $x \in \mathbb{R}^n$ the sales of a product category, and $y \in \mathbb{R}^n$ the "final" values of wILI reported by CDC, where n is the number of time steps (months or weeks) in the datasets. Then analysis based on the correlation criterion

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

is probably the simplest approach to this problem. Intuitively, one wants to find features that are highly correlated with the response variable and thus make a good predictor. In our influenza prediction problem, we want to find and select sales data of products that are most related to influenza, more specifically, to the final values of wILI. For terminology convenience, we refer the CDC-reported final values of wILI as stable wILI and initial versions of wILI as unstable wILI. In its simplest form, correlation coefficient (1) between sales of each product and stable wILI is computed and then sorted from high to low by the magnitude. The criteria for feature selection is either defined by a given k for the number of features desired or specified by a cut-off on the magnitude of the correlation.

One major drawback of this correlation based selection is that it ignores the correlation between features (between sales data of products themselves, as well as between the sales data and the data sources that are currently being used the Nowcast). Since our goal is to improve upon the existing Nowcast by *adding* sales data of products (while keeping the data already being used), we can look at correlation between residual of the Nowcast and sales of a product. I.e. in (1) above, while x still represents sales of a product, we let $y = \mathrm{wILI} - \mathrm{Nowcast}^0$, where $\mathrm{Nowcast}^0$ is a vector of estimates made by the existing Nowcast. In a sense, we want to find features that explain away the error between the ground truth (stable wILI) and the current estimates.

5.2 Orthogonal Matching Pursuit

In the field of signal processing, the Orthogonal Matching Pursuit (OMP) is an algorithm for the recovery of a high-dimensional sparse signal based on a small number of noisy measurements.(Cai & Wang, 2011) It aims to approximate the solution of one of the following problems (Rubinstein, Zibulevsky, & Elad, 2008), the sparsity constrained problem,

$$\underset{\beta}{\text{minimize}} \ \|y - X\beta\|_2^2 \qquad \text{subject to} \quad \|\beta\|_0 \leq K, \tag{2}$$

and the error constrained problem

$$\underset{\beta}{\text{minimize}} \ \|\beta\|_0 \qquad \text{subject to} \quad \|y - X\beta\|_2^2 \leq \epsilon, \tag{3}$$

where $\|\beta\|_0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$. In our influenza nowcasting problem, $y \in \mathbb{R}^n$ is final values of wILI and $X \in \mathbb{R}^{n \times p}$ is sales data, with p being the number of product categories in the dataset. Note that (2) is exactly the definition of the best subset selection problem (Efroymson, 1966; Draper & Smith, 1966).

The above two problems are non-convex (due to the ℓ_0 pseudo-norm of the minimization variable) and have shown to be NP-hard (Natarajan, 1995). The OMP is a greedy, iterative algorithm that at each step selects from the rest features the one that is most correlated with the current residual, which is the residual of projecting wILI onto linear space spanned by the already selected subset of features. In more detail, it can be stated as the following:

Algorithm 1: Orthogonal Matching Pursuit (OMP) algorithm

1 function OMP (X, y);
Input : Feature variables X (sales + SAR3) and response variable y (wILI)
Output : A sparse subset of features S selected by OMP
2 Normalization: normalize X such that each column $\|x_d\|_2 = 1$;
3 Initialization: set $S = \emptyset, r = y$;
4 **while** *stopping criterion not met* **do**
5 Find the variable x_d that solves the optimization problem $\max_d |x_d^T r|$;
6 Add the variable x_d to the set of selected variables, update $S = S \cup \{x_d\}$;
7 Compute the projection of y onto the linear space spanned by the elements in S, $Py = S(S^T S)^{-1} S^T y$, update residual $r = y - Py$;
8 **end**

The stopping criterion in Line 3 of Algorithm 1 can be a fixed number of iterations (thus the number of features included in the end), or based on minimizing the cross-validated mean sqaured error (MSE), or other more dedicated conditions (Cai & Wang, 2011).

5.3 LASSO

Another commonly used algorithm for sparse feature selection is the LASSO algorithm (Tibshirani, 1996). The LASSO can be seen as a convex relaxation of the best subset selection problem (2) by relaxing the ℓ_0 constraint to the ℓ_1 constraint,

$$\underset{\beta}{\text{minimize}} \ \|y - X\beta\|_2^2 \qquad \text{subject to} \quad \|\beta\|_1 \leq s, \tag{4}$$

or equivalently,

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{5}$$

where $\lambda > 0$ is a regularization parameter.

For a given λ, solution to (5) can be effectively computed through various convex optimization methods, such as proximal gradient descent, accelerated proximal gradient descent, alternating direction method of multipliers (ADMM), and coordinate descent. The magnitude of λ makes a direct impact on the sparsity of the solution. By varying the parameter λ between ∞ and 0, a path of the solution, known as Lasso path, can be obtained. By plotting out the path, we can get a direct sense of what features are going to be included in the solution as we gradually relax the sparsity constraints (by decreasing λ).

The final value of λ used for feature selection is often determined by minimizing the cross-validated MSE,

$$\hat{\lambda} = \arg\min_{\lambda} \mathrm{CV}(\lambda).$$

Another way to choose λ is based on the one standard error rule. In this case, we first find the usual minimizer $\hat{\lambda}$ as above, and then increase λ as much as we can such that the resulting cross-validation error is still within one standard error of $\mathrm{CV}(\hat{\lambda})$, i.e.

$$\mathrm{CV}(\lambda) \leq \mathrm{CV}(\hat{\lambda}) + \mathrm{SE}(\hat{\lambda}).$$

5.4 Heurisitc Selection Based on Weighted Linear Regression

All the methods discussed so far treat data points at different time steps uniformly and as independent observations. This assumption is often violated with time-series data. In many cases, data from neighboring time points are highly correlated. Coming to our problem, it is almost sure that the flu prevalence this week is highly correlated with that in last week, since infection with and recovery from the disease is not instantaneous and transmission among population often makes the trend continue. Also because the influenza is seasonal, one can imagine the influenza activity at this time is somehow similar to those at around the same time in the past few years. Finally, all the above models assume we have and use the "true" label — final values of wILI. Due to backfilling effect discussed in the Introduction section 1, it means we cannot use the most recent data for feature selection without violating the assumption. This can be quite a disadvantage since the most recent wILI data, despite being subject to backfilling, still provides valuable information about the influenza prevalence at present. To alleviate these issues, (Farrow, 2016; Farrow & Rosenfeld, 2017) designed the following weights for use in weighted linear regression (WLR),

$$\begin{aligned} a &= \tfrac{1}{20} + \tfrac{19}{20}\exp(-(\min(dw\%w_0, w_0 - dw\%w_0)/2)^2), \\ b &= 2^{-(dw/w_0)}, \\ c &= 1 - 2^{-(dw/1)}, \\ \text{weight}(dw) &= a \cdot b \cdot c, \end{aligned} \tag{6}$$

where dw is the number of weeks between current week and a past week, $w_0 = 52.2$ is the average number of weeks in a year, $\%$ is taking the modulo. The factor a considers the seasonal effect of influenza, factor b tries to capture correlation with influenza in the past few weeks by assigning relatively more weights to them, and factor c is meant to "penalize" the most recent wILI data as they are subject to revision.

We come up with an empirical, heuristic way to select features based on this method (HSWLR) for data at the HHS regional level, currently the finest geographical for CDC reported wILI. Data at finer spatial granularities (e.g. states) can be aggregated to the regional level. We perform the WLR above with two variates, SAR3 and sales of a product category, against wILI, and compare with the SAR3 estimate itself, for each of the product categories in the dataset. More specifically, at a past week t in the dataset for a region, we do a weighted linear regression with all the sales data of a product and SAR3 data *up to* week t, and the wILI report published on week t (which contains unstable wILIs for weeks $t-1, t-2, \ldots$, and stable wILIs for weeks months earlier than week t), and obtain an estimate of influenza prevalence at week t for that region. We evaluate the error with respect to the final wILI for that week and region. The above steps are repeated for each week in the dataset and the errors from these weeks are summed, squared, and averaged, to obtain the MSE of this bivariate WLR for a product, for each region separately. Note that the WLR is not performed once, but rather at every week for every region using all the data up to that week for that region. We then compare this MSE with the MSE of the SAR3 estimate for each region and count the number of regions where the MSE from the bivariate WLR is smaller than that from the SAR3 (thus an improvement).

We can do this procedure for each product, and rank them by the number of regions it improves compared to the SAR3 itself. Heuristic feature selection can then be made by either selecting the top k features in this ranking, or setting a threshold on the number of regions it should improve.

5.5 Nowcast-based Sequential Feature Selection

So far we have assumed a linear model

$$y = X\beta + \epsilon, \tag{7}$$

where y is wILI and X is sales data of product categories (plus SAR3 for some cases).

We have also implicitly made the assumption that the number of observations are equal for all variables (features and label). In reality, this is hardly true, as different data sources can start and end at different times or be missing during some times in the middle. What makes the situation more complicated is that the data sources can be of different spatial granularities: some are national, others by HHS regions, others by states, etc. The Nowcast provides a nice way to deal with these problems. Readers can refer to (Farrow, 2016; Farrow & Rosenfeld, 2017) for details. Essentially, it trains each data source into a sensor by the weighted linear regression using weights (6). Then

covariance matrix of errors between each sensor and measurements (wILI) is computed when the sensor and measurements are simultaneously available. Sensor fusion can then be performed based on this covariance matrix.

The capability of the Nowcast system to easily accommodate missing and spatially heterogeneous data makes it powerful but also makes it distinct from and less intuitive than the traditional linear regression model (7). It also raises the question that whether features selected by the previous methods (which is based on (7)) will still make significant improvement when added to the final Nowcast model.

For this we consider the Nowcast-based sequential feature selection algorithm (NSFS). The sequential feature selection (SFS) is similar to the OMP in that both are greedy, iterative, forward selection algorithms. But unlike the OMP that at each step it selects a feature that minimizes the quadratic loss of the linear model (7), the SFS in principle works for any any model and objective function as long as they are defined. At each step, it computes value of the objective function for a given model for each of the features and select the feature that minimizes the objective. In the problem of influenza nowcasting, the model is the whole Nowcast process. I.e. for each feature (sales of a product), we train it into a sensor using weighted linear regression (single variate, with weights (6)), temporarily incorporate the sensor into the Nowcast system and evaluate performance of the resulting Nowcast. The performance evaluation will be discussed in more detail in Section 5.6. Briefly, it is the MSE of doing Nowcast at each of the past time steps (weeks) in the dataset using all the information up to that week. We use this MSE as the objective in NSFS, and select from rest features the one that optimizes the objective and officially incorporate it into the Nowcast, and then continue for the rest of the features.

One major drawback of the SFS is that if the model and objective function is complicated and takes much time to evaluate, then the time cost is expensive. As in our Nowcast-based sequential feature selection (NSFS), the time complexity is $\Theta(np^2T)$ for p product categories (columns) and n time steps (rows) in the dataset, where T is the number of operations to train a sensor (using WLR with weights (6)) and run a Nowcast with the sensor included and hence depends on the number of weeks included (which can be significantly larger than n as some other data sources in the Nowcast have been available since much earlier) and the number of features in the Nowcast. Additionally, if we want to do it for each region separately, the above time cost needs to be multiplied by a factor of 10. Another drawback of the NSFS is that, being a greedy algorithm, it may converge to a local optimum rather than global (same is true for the OMP).

5.6 Evaluation

After a subset of features is selected, we perform evaluation by doing nowcast at each of the past weeks in the dataset sequentially, and computing the error with respect to the final values of wILI for that week. Then the errors from all the weeks are squared, summed, and averaged to obtain the final objective,

$$\text{MSE} = \frac{1}{N}\sum_{i=t_1}^{t_n}(\text{Nowcast}(i) - \text{wILI}(i))^2. \tag{8}$$

This is essentially what would be in reality: to nowcast the influenza at week t, we use all the information available up to that week (e.g. initial version of wILI for week $t-1$ and data of other sources at week t if available). The final objective is the mean of the squared errors for all the weeks.

We point out that while the metric above provides a reasonable way for evaluating performance of selected features, it is not a perfectly fair way to compare between different feature selection methods. This is because during feature selection we selected a subset of features using data from all the time steps (weeks) in the dataset (even the NSFS above is using the final MSE (8) to select features). But in the evaluation above, we are using this "final" subset of feature for each of the past weeks. In other words, to compare more fairly between different feature selection methods, one would perform feature selection dynamically at each of the past weeks using information up to that week. This means the features selected by a feature selection method may vary from week to week as time progresses.

On the other hand, since the main goal of this work is to select a good subset of features, the evaluation metric should still provide helpful information regarding the "goodness" of features.

6 Results

We apply each of the above methods to the data described in Section 4. Whenever available (e.g. the second and third datasets, *Flu-117* and *Priority-List*), we use the sales data in units instead of volume as it is more intuitive.

6.1 Correlation with wILI

We compute the correlation of sales of each product with stable wILI and sort them by the magnitude of the correlation in descending order. Table 1 shows the results for the first few entries after applying (1) to the first dataset (*Full-categories*).

Table 1: Correlation of sales of each product with stable wILI for dataset *Full-categories*

Product	Correlation with wILI
PERSONAL THERMOMETERS	0.981085
INTERNAL ANALGESIC LIQUIDS	0.971097
GASTROINTESTINAL - LIQUID	0.917352
FZ PIZZA CRUSTS/DOUGH	0.916530
NASAL STRIPS	0.912657

Two observations from Table 1: (i) the sales data of the listed products show very high correlation with wILI ($\rho > 0.91$); (ii) product like "FZ PIZZA CRUSTS/DOUGH" shows a correlation with wILI as high as 0.9165, although it is very unlikely by common sense. Both of these observations are likely caused by the relatively few data points for each product in this dataset (one year monthly, national data, 13 points). Indeed, as an example, the correlation of "NASAL STRIPS" dropped to 0.6779 in the second dataset *Flu-117* simply due to more data (weekly, national data from 2012 - 2017, 261 points).

The fact that flu activity is high in the winter and low in the summer implies that any product whose sales coincides with such a seasonal pattern is likely to show high correlation with the flu, regardless whether the two are actually correlated. This is especially problematic when we have few data (e.g. one flu season for the first dataset) and can lead to false discoveries, which is probably the case for FZ PIZZA CRUSTS/DOUGH. For this reason, we put our main focus on the latter two datasets. Besides using more data, another simple way to mitigate the artificial effect on correlation introduced by seasonal coincidence is to consider in-season data only. In-season period is defined from week 40 (early October) of a year to week 20 (early May) of next year, i.e. when influenza activity is significant. Focusing on in-season influenza activity also agrees with our main interest.

6.2 Features Selected by the OMP

We apply the OMP to our second dataset, *Flu-117*. Since the goal is to improve upon what can already be done, we have included SAR3, which makes the major contribution to the current Nowcast, as one feature, along with sales of product categories when performing the OMP. The first six features that are selected are (in sequence)

> SAR3, SLEEPING REMEDIES, INTERNAL ANALGESIC LIQUIDS, BUTTER/BUTTER BLENDS, DISPOSABLE DIAPER, LIQUID VITAMINS/MINERALS.

The SAR3 is selected in the first iteration indicates that wILIs in the recent past, although being subject to backfilling, are still very useful and a better feature than sales of any product category.

When using the cross validation mean square error (CV-MSE) as the model selection criterion, four features are selected (using ten-fold cross-validation),

> SAR3, SLEEPING REMEDIES, INTERNAL ANALGESIC LIQUIDS, BUTTER/BUTTER BLENDS.

We have also performed the OMP on the third dataset *Priority-List*. As mentioned, this set contains weekly sales data of 21 categories of products from 2012 to 2017 for 45 states in the US. Since

the wILI of the finest geographical resolution currently available is at the HHS regional level, we have aggregated the data by the HHS regions they belong to. Then the OMP with cross validation (OMP-CV) is performed for each HHS region separately. Table 2 shows the number of features selected (including SAR3) for each region. Details on the selected features for each region are shown in Table 6 in the Appendix. From the results there we note that the SAR3 is selected for each of the ten regions by the algorithm. Aside from that, the number and kinds of other features selected vary from region to region. This can make sense because shopping preferences of people, popularity of the product and relative easiness in getting access to the product can all vary from region to region.

Table 2: Number of features selected by the OMP-CV for each region from dataset *Priority-List*.

HHS1	HHS2	HHS3	HHS4	HHS5	HHS6	HHS7	HHS8	HHS9	HHS10
2	1	5	3	3	3	4	3	4	2

One interesting finding is that only one feature (SAR3) is selected for HHS2 region, which means that the cross-validation decides that it is not worth taking in another feature in terms of out-of-sample prediction error.

6.3 Sparse Representation through the LASSO

The subset of features selected by the LASSO is very different from that by the OMP when applied to the same dataset *Flu-117*. Most notably, thirty two features are selected by the LASSO with (ten-fold) cross validation, compared to only four by the OMP. Ordered by the magnitude (absolute value) of the coefficient, the first ten of the thirty two features selected are

> PERSONAL THERMOMETERS, SAR3, INTERNAL ANALGESIC LIQUIDS, RFG YOGURT DRINKS, SS HONEY, FZ BLENDED FRUIT JUICE CONCENTRATE, HAND SANITIZERS, LIQUID VITAMINS/MINERALS, SS HONEY, BLOOD PRESSURE KIT

We have also looked at the solution path of the LASSO applied to this dataset, as shown in Figure 2. Interestingly, when the regularization parameter λ is big (left side of the figure), the solution appears unstable. Some features that are initially selected by the LASSO vanishes from the solution later as λ decreases. This can be related to the fact that data of some entries in the datasets are highly correlated (e.g. data for a product category and for a subcategory, and the subcategory is the major subcategory of that category).

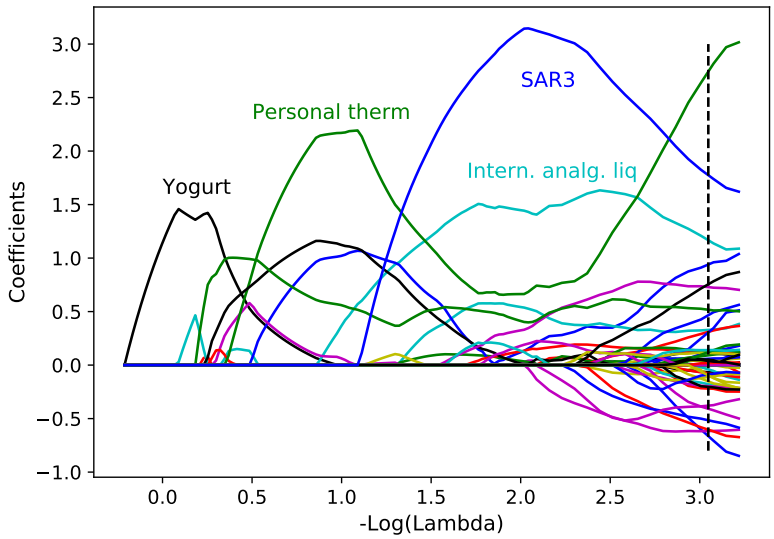


Figure 2: Solution path of the LASSO applied to the dataset *Flu-117*. The black dashed line locates the `Lambda` with minimum cross-validation error.

Similar to the OMP, when applied to the dataset *Priority-List*, the SAR3 is selected by the LASSO with cross validation (LASSO-CV) for all the HHS regions, and the number and kinds of other features selected vary from region to region. Table 3 lists the number of features selected by the LASSO. The selected features for each region are listed in Table 7 in the Appendix.

Table 3: Number of features selected by the LASSO-CV for each region from dataset *Priority-List*.

HHS1	HHS2	HHS3	HHS4	HHS5	HHS6	HHS7	HHS8	HHS9	HHS10
13	4	20	14	13	8	15	6	18	12

As is for the dataset *Flu-117*, the LASSO-CV selects more features than the OMP-CV for each region. This is generally true for selections based on the shrinkage method (such as the LASSO). The LASSO shrinks the magnitude of all the coefficients by translating the coefficients towards zeros by a constant value and sets them to zero if they reach it (known as soft-thresholding). This means those supposedly large coefficients are also shrunk. To minimize the CV-error, more features tend to be included. That is why the heuristic, one standard error rule is introduced (discussed in Section 5.3), where it tries to select fewer features with larger λ while still being close to the minimum of the CV-error.

6.4 Heuristic Selection based on Weighted Linear Regression

As discussed in the Section 5.4, we perform bivariate (SAR3 + sales of a product), weighted linear regression with weights specified in (6), and compute the out-of-sample MSE. Figure 3 shows an example of doing such regression for one of the product categories in the dataset *Priority-List*.

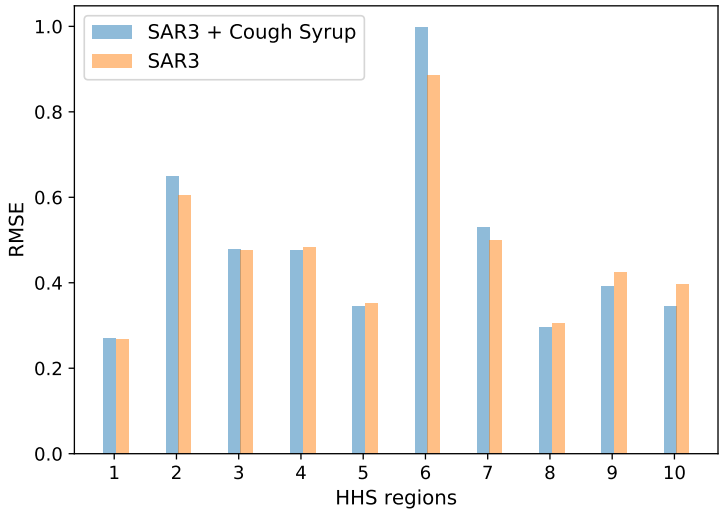


Figure 3: Out-of-sample root mean squared error (RMSE) of weighted linear regression using SAR3 and sales of Cough Syrup as covariates for each region (blue), compared to that of the SAR3 itself.

From Figure 3, we see that when the sales data of Cough Syrup is added, the combined predictor improves for some of the regions (4, 5, 8, 9, 10) while deteriorates for others. The additional data source does not make improvement on all regions may appear surprising. But again this is because when we do the weighted linear regression up to week t, we use all information available by that week, including unstable values of wILI for weeks $t-1, t-2, \ldots$. But the evaluation is performed using the final, stable values of wILI for week t.

We use the heuristic feature selection described in the HSWLR method 5.4, and set the threshold on the number of regions the bivariate WLR should improve over the SAR3 to be three. Table 4 shows features selected from the dataset *Priority-List*, along with the number of regions it improves.

Table 4: Features selected by the HSWLR, along with the number of regions it improves when combined with SAR3 over SAR3 alone.

PERSONAL THERMOMETERS	9	COLD_ALLERGY_SINUS LIQUID	5
INTERNAL ANALGESIC LIQUIDS	8	SORE THROAT REMEDY LIQUIDS	4
COUGH SYRUP	5	COLD_ALLERGY_SINUS TABLETS	3

6.5 Nowcast-based Sequential Feature Selection

We have employed the NSFS for the dataset *Priority-List*. Due to time cost, we have made the assumption that the features selected by NSFS are the same across all the regions, leading to a reduction of factor 10 in the run time.

In Figure 4, we plot the objective (sum of out-of-sample MSEs of ten regions) as a function of the algorithm's iteration number which is equal to the number of additional features included.

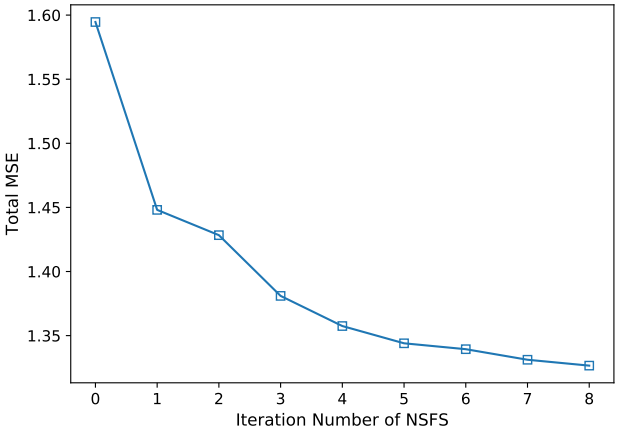


Figure 4: Total MSE (sum of MSEs of ten regions) as a function of the iteration number of the NSFS.

The point at iteration number 0 corresponds to the case when no sales data is included, i.e. performance of the current Nowcast. We see that the total MSE declines sharply in the first few iterations, and more slowly as more features are selected.

Heuristically we select eight features in consideration of both the time cost and the trend of the error curve in Figure 4. The eight features selected are (in sequence of the iteration):

> INTERNAL ANALGESIC LIQUIDS, PERSONAL THERMOMETERS, SLEEPING AID TABLETS, SS BOTTLED ORANGE JUICE, SLEEPING AID LIQUIDS, COUGH SYRUP, SORE THROAT REMEDY LIQUIDS, INTERNAL ANALGESIC TABLETS.

6.6 Evaluation of Features Selected by Different Methods

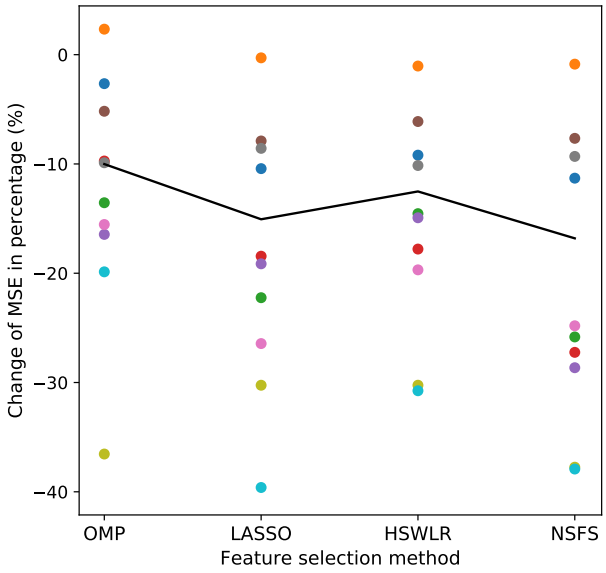
We evaluate the different subsets of features selected by the different algorithms above using the method introduced in 5.6. The evaluation is performed only on features from the third dataset *Priority-List*, where weekly, regional (by aggregating states) sales data is available.

Table 5 summarize the MSEs (8) after adding the features selected to the Nowcast for each feature selection method, along with the MSE of the original Nowcast (NC0).

Figure 5 show changes of MSE relative to the original Nowcast, in percentage, for each region (colored solid dots) as well as for the average of all the ten regions (black line). A negative number indicates a reduction of the MSE, thus an improvement of the performance. We see that features selected by all methods improve the Nowcast for almost all the regions, when added to the Nowcast system. Features selected by different selection methods demonstrate different impacts on different regions. Overall, the LASSO and NSFS seem to work best for our Nowcast problem, achieving a reduction of 15.06% and 16.81% respectively in the total MSE relative to the original Nowcast.

Table 5: MSEs after adding features selected by different algorithms to the Nowcast for each HHS region, using evaluation method discussed in 5.6. Note entries of "NC0" column are the MSEs of the original Nowcast, i.e. without no sales data included.

	NC0	OMP	LASSO	HSWLR	NSFS
HHS1	0.0447	0.0435	0.0400	0.0405	0.0396
HHS2	0.2829	0.2895	0.2821	0.2800	0.2805
HHS3	0.1201	0.1038	0.0934	0.1026	0.0891
HHS4	0.1264	0.1141	0.1031	0.1039	0.0919
HHS5	0.0683	0.0571	0.0553	0.0581	0.0488
HHS6	0.4871	0.4618	0.4486	0.4573	0.4498
HHS7	0.1533	0.1294	0.1127	0.1231	0.1152
HHS8	0.0625	0.0563	0.0572	0.0562	0.0567
HHS9	0.1227	0.0779	0.0856	0.0856	0.0764
HHS10	0.1266	0.1015	0.0765	0.0877	0.0786



Method	ΔMSE (%)
OMP	-10.01
LASSO	-15.06
HSWLR	-12.52
NSFS	-16.81

Figure 5: Left: Change of MSE relative to the original Nowcast when sales data selected by different methods are added to the Nowcast separately. The colored solid dots are changes of MSE in each region and the black line shows the average. Right: values of the average change of MSE for each method.

7 Discussion

The evaluation results in Section 6.6 indicate that the LASSO and the NSFS work the best for our problem in that maximum reduction in the test MSE is achieved when the features selected by them are added to the Nowcast. The reason that the NSFS works better than other feature selection methods is quite straightforward. As discussed in Section 5.5, the NSFF works directly on the Nowcast model and the final evaluation metric. The Nowcast model is in general not equivalent to the linear model (7) assumed by other feature selection methods. The relatively poor performance of the HSWLR is probably due to the selection procedure itself besides assuming the linear model. It selects features (sales of products) with the hope that they make improvement on as many regions as possible, while ignoring how much it improves (in terms of reduction of MSE) for each region. Also it is one-iteration approach, heuristically selecting k features based on the criterion just described, instead of iterative approach employed by others.

For the OMP and the LASSO results reported in this work, we have used the 10-fold cross-validation with randomly split folds, for model selection. We realize that this is not ideal as the random splitting can jumble up the time aspect and does not reflect training/testing splits like we would see in real application for time series data. Improvement will be made in the near feature regarding this aspect. At the same time, we believe this random splitting in cross-validation has a greater impact on iterative, greedy approaches than on convex, non-greedy algorithms, and is the reason the LASSO performs significantly better than the OMP for our problem.

In addition to improving formulation of the cross-validation, another work that we are working on right now is to use to the one standard error rule instead of the usual min-CV rule (as in this work) for choosing the regularization parameter in the LASSO (5.3). This would lead to a model with a smaller number of features and potentially better interpretability.

Finally, we comment that the results in Figure 5 should be interpreted more qualitatively than quantitatively. As we have discussed, it is not perfectly fair to compare between feature selection methods using the evaluation method in Section 5.6. This is because the current methods selected features based on the full data set. In the evaluation step, we go back and evaluate the Nowcast in a rolling fashion over time, including features selected with knowledge of future data. To compare between different methods more fairly, and to obtain truly out-of-sample test errors for selected features, we need to perform feature selection dynamically, for each past week, using all information up to that week to select features and do the Nowcast and evaluation. This is planned in the next step of our work. On the other hand, results in Section 6 should give an idea of the goodness of these features. In particular, by comparing the features selected by the different methods, we see that product categories like INTERNAL ANALGESIC LIQUIDS, PERSONAL THERMOMETERS, COUGH SYRUP, and SLEEPING AID LIQUIDS are selected by most of the methods. We think these features will help the Nowcast in the future and should be included if the data for them is available timely.

8 Conclusion

In this work, we have explored a few popular feature selection algorithms and applied them to the problem of influenza nowcasting in the Unites States on the datasets of sales of grocery products. All algorithms considered select features that help improve the accuracy of the current Nowcast system, with the LASSO and the NSFS achieving the maximum reduction in the errors. When applied to individual HHS region separately, we find the number and kinds of features selected generally vary from region to region. For the influenza nowcasting problem, the features selected by the OMP are sparser than those selected by the LASSO, both for each region and at the national level. Due to time cost of the NSFS, we make the assumption that the features selected are the same for all the regions. It is expected that if the features selected are allowed to vary for each region in the NSFS, the reduction in the errors may be even greater. We have also discussed limitations in the current results and improvements planned for the future.

9 Acknowledgment

First and foremost, I want to thank my DAP advisor, Prof. Roni Rosenfeld for giving me the opportunity to work on this project. Roni has been extremely helpful and patient in guiding me through this work, which I have enjoyed and learned a lot over the past year. I am very grateful for my other committee member, Prof. Ryan Tibshirani, for a lot of insightful comments and suggestions along the project. I am also thankful for discussions with other members in the Delphi group, including David Farrow, Justin Hyun, Logan Brooks, and Lisheng Gao.

References

Beale, E. M. L., Kendall, M. G., & Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, *54*(3/4), 357-366.

Brauer, F., Castillo-Chavez, C., & Castillo-Chavez, C. (2001). *Mathematical models in population biology and epidemiology* (Vol. 1). Springer.

Brooks, L. C., Farrow, D. C., Hyun, S., Tibshirani, R. J., & Rosenfeld, R. (2015, 08). Flexible modeling of epidemics with an empirical bayes framework. *PLoS Comput Biol*, *11*(8), e1004382.

Cai, T. T., & Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information theory*, *57*(7), 4680–4688.

CDC. (2016). Overview of influenza surveillance in the united states. *Centers for Disease Control and Prevention, Atlanta, Georgia, U.S. https://www.cdc.gov/flu/weekly/overview.htm*.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, *43*(1), 129–159.

Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J., & Vespignani, A. (2007). Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Med*, *4*(1), e13.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis*. Wiley.

Efroymson, M. (1966). Stepwise regression — a backward and forward look. *Eastern Regional Meetings of the Institute of Mathematical Statistics*.

Farrow, D. (2016). *Modeling the past, present, and future of influenza* (Unpublished doctoral dissertation). Carnegie Mellon University. (Available online at https://delphi.midas.cs.cmu.edu/ dfarrow/thesis.pdf)

Farrow, D., & Rosenfeld, R. (2017). Multiple resolution nowcasting of influenza through sensor fusion. *In preparation*.

Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006, 7). Strategies for mitigating an influenza pandemic. *Nature*, *442*, 448-452.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning — data mining, inference, and prediction* (2nd ed.). Springer-Verlag New York. (Available online at https://web.stanford.edu/ hastie/Papers/ESLII.pdf)

Hocking, R. R., & Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, *9*(4), 531-540.

Molinari, N., Ortega-Sanchez, I., Messonnier, M., Thompson, W., Wortley, P., Weintraub, E., & Bridges, C. (2007). The annual impact of seasonal influenza in the us: measuring disease burden and costs. *Vaccine*, *25*(27), 5086–5096.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM journal on computing*, *24*(2), 227–234.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E*, *66*(1), 016128.

Rubinstein, R., Zibulevsky, M., & Elad, M. (2008). Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit. *Cs Technion*, *40*(8), 1–15.

Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, *109*(50), 20425–20430.

Shumway, R. H., & Stoffer, D. S. (2013). *Time series analysis and its applications*. Springer Science & Business Media.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

WHO. (2016). Influenza (seasonal). fact sheet no. 211. *World Health Organization, Geneva, Switzerland. http://www.who.int/mediacentre/factsheets/fs211/en/*.

Appendix

Table 6 lists features selected by the OMP-CV (in alphabetical order) on the dataset *Priority-List* for each of the ten regions.

Table 7 lists features selected by the LASSO-CV (in alphabetical order) on the dataset *Priority-List* for each of the ten regions.

Table 6: Features selected by cross-validated OMP for each region

HHS1	PERSONAL THERMOMETERS, SAR3
HHS2	SAR3
HHS3	ANTACID LIQUID POWDER, COLD ALLERGY SINUS TABLETS, INTERNAL ANALGESIC LIQUIDS, SAR3, SLEEPING AID TABLETS
HHS4	ANTACID LIQUID POWDER, INTERNAL ANALGESIC LIQUIDS, SAR3
HHS5	INTERNAL ANALGESIC LIQUIDS, SAR3, SLEEPING AID LIQUIDS
HHS6	ANTACID LIQUID POWDER, INTERNAL ANALGESIC LIQUIDS, SAR3
HHS7	INTERNAL ANALGESIC LIQUIDS, SAR3, SLEEPING AID LIQUIDS, SS BOTTLED TOMATO VEGETABLE
HHS8	HAND SANITIZERS, SAR3, SORE THROAT REMEDY LIQUIDS
HHS9	COLD ALLERGY SINUS TABLETS, HAND SANITIZERS, PERSONAL THERMOMETERS, SAR3, SLEEPING AID TABLETS
HHS10	COLD ALLERGY SINUS TABLETS, SAR3

Table 7: Features selected by cross-validated LASSO for each region

HHS1	ANTACID LIQUID POWDER, COUGH SORE THROAT DROP, HAND SANITIZERS, INTERNAL ANALGESIC LIQUIDS, NASAL SPRAY DROPS INHALER, NASAL STRIPS, PERSONAL THERMOMETERS, SAR3, SLEEPING AID TABLETS, SS BOTTLED APPLE JUICE, SS BOTTLED GRAPE JUICE, SS BOTTLED ORANGE JUICE, SS BOTTLED TOMATO VEGETABLE
HHS2	COLD ALLERGY SINUS TABLETS, INTERNAL ANALGESIC LIQUIDS, PERSONAL THERMOMETERS, SAR3
HHS3	ANTACID LIQUID POWDER, COLD ALLERGY SINUS TABLETS, COUGH SYRUP, COUGH SORE THROAT DROP, FACIAL TISSUE, HAND SANITIZERS, INTERNAL ANALGESIC LIQUIDS, NASAL ASPIRATORS, NASAL SPRAY DROPS INHALER, NASAL STRIPS, PERSONAL THERMOMETERS, RFG ORANGE JUICE, SLEEPING AID LIQUIDS, SAR3, SLEEPING AID TABLETS, SORE THROAT REMEDY LIQUIDS, SS BOTTLED APPLE JUICE, SS BOTTLED GRAPE JUICE, SS BOTTLED ORANGE JUICE, SS BOTTLED TOMATO VEGETABLE
HHS4	ANTACID LIQUID POWDER, COUGH SORE THROAT DROP, FACIAL TISSUE, HAND SANITIZERS, INTERNAL ANALGESIC LIQUIDS, INTERNAL ANALGESIC TABLETS, NASAL SPRAY DROPS INHALER, NASAL STRIPS, PERSONAL THERMOMETERS, SAR3, SLEEPING AID LIQUIDS, SS BOTTLED GRAPE JUICE, SS BOTTLED ORANGE JUICE, SS BOTTLED TOMATO VEGETABLE
HHS5	ANTACID LIQUID POWDER, COLD ALLERGY SINUS TABLETS, COUGH SORE THROAT DROP, FACIAL TISSUE, INTERNAL ANALGESIC LIQUIDS, INTERNAL ANALGESIC TABLETS, NASAL ASPIRATORS, NASAL STRIPS, PERSONAL THERMOMETERS, SAR3, SLEEPING AID LIQUIDS, SS BOTTLED ORANGE JUICE, SS BOTTLED TOMATO VEGETABLE
HHS6	ANTACID LIQUID POWDER, COLD ALLERGY SINUS LIQUID, COLD ALLERGY SINUS TABLETS, HAND SANITIZERS, INTERNAL ANALGESIC LIQUIDS, NASAL ASPIRATORS, PERSONAL THERMOMETERS, SAR3
HHS7	ANTACID LIQUID POWDER, COLD ALLERGY SINUS TABLETS, COUGH SORE THROAT DROP, HAND SANITIZERS, INTERNAL ANALGESIC LIQUIDS, NASAL STRIPS, PERSONAL THERMOMETERS, RFG ORANGE JUICE, SAR3, SLEEPING AID LIQUIDS, SLEEPING AID TABLETS, SS BOTTLED APPLE JUICE, SS BOTTLED GRAPE JUICE, SS BOTTLED ORANGE JUICE, SS BOTTLED TOMATO VEGETABLE
HHS8	INTERNAL ANALGESIC LIQUIDS, PERSONAL THERMOMETERS, RFG ORANGE JUICE, SAR3, SLEEPING AID LIQUIDS, SORE THROAT REMEDY LIQUIDS
HHS9	ANTACID LIQUID POWDER, COLD ALLERGY SINUS TABLETS, COUGH SYRUP, FACIAL TISSUE, HAND SANITIZERS, INTERNAL ANALGESIC LIQUIDS, INTERNAL ANALGESIC TABLETS, NASAL ASPIRATORS, NASAL SPRAY DROPS INHALER, PERSONAL THERMOMETERS, SAR3, SLEEPING AID LIQUIDS, SLEEPING AID TABLETS, SORE THROAT REMEDY LIQUIDS, SS BOTTLED APPLE JUICE, SS BOTTLED GRAPE JUICE, SS BOTTLED ORANGE JUICE, SS BOTTLED TOMATO VEGETABLE
HHS10	COLD ALLERGY SINUS TABLETS, COUGH SYRUP, COUGH SORE THROAT DROP, FACIAL TISSUE, HAND SANITIZERS, INTERNAL ANALGESIC LIQUIDS, PERSONAL THERMOMETERS, SAR3, SLEEPING AID LIQUIDS, SLEEPING AID TABLETS, SS BOTTLED APPLE JUICE, SS BOTTLED ORANGE JUICE