COMP 551 Project 1 Write-up

Anny HANG, Sameen MAHTAB, Michelle WANG

October 21, 2020

Abstract

We used supervised K-nearest neighbors and decision trees to predict COVID-19 hospitalization cases given data on search trends for various symptoms. One dataset consisted of search trends for several symptoms for regions in the United States and the other consisted of data about individuals who got COVID-19 including the number of new cases of hospitalizations. First, we processed the datasets and merged them together. Then we visualized and clustered the data to acquire a better understanding. After that we analyzed the results of the two models running on the merged dataset.

1 Introduction

Google Trends has been a topic of discussion for health care research (1, 2). With the number of new cases of COVID-19 oscillating, finding a way to predict new cases is important to control outbreaks. For this project, we visualized the search trends of popularity of COVID-19 related symptoms across different regions, and we used supervised machine learning algorithms to predict new cases of hospitalization based on search trends of symptoms.

We worked with two datasets. The first dataset included information about weekly patterns of search trends of symptoms related to COVID-19 (3). The second dataset included a compilation of hospitalization cases from different regions of the world (4). Information of search trends of symptoms and new hospitalization cases were merged together by region and by date in order to create the data needed for our analysis. We visualized our merged dataset using Principal Component Analysis (PCA), and performed clustering using a K-means model on PCA-reduced and non-reduced data. To predict new hospitalization cases from patterns related to search trends of symptoms, we used two regression models, K-nearest neighbours (KNN) and decision trees, to investigate their performance given a training set based on region and another training set based on date. We found that splitting the data by date generates a more accurate KNN model and decision tree model.

2 Datasets

The first dataset we used contained information about the relative popularity of 422 search topics (symptoms) in 16 different regions (states) in the United States, at a weekly resolution (from the week of January 6th, 2020 to that of September 28th, 2020). The second one contained new daily hospitalization cases for 424 regions (including regions that are outside the United States) from December 31st, 2019 to October 8th, 2020.

For the search trends dataset, we first filtered the symptoms by removing columns that consisted entirely of NaNs: this left us with 121 symptoms. We considered removing symptoms that had a low frequency of non-NaN values, but ultimately decided against this because the provider of this dataset used NaN both when the value was missing and when it was 0. Our reasoning was that even symptoms that had a lot of NaNs/0s could potentially be useful, for example if their non-NaN values correlated with the number of new hospitalizations.

We then counted the number of non-NaN values across symptoms for each region and decided to remove 4 regions that had less than 500 non-NaN values. Please see the Appendix for plots that we used to help use determine filtering criteria (thresholds) in our preprocessing steps.

For the hospitalization dataset, we converted the data to a weekly format that matched the search trends dataset's resolution. We then removed all rows that had a NaN value in the "hospitalized_new" column. Then, we changed all negative hospitalization values to 0 (since we did not care about hospitalization discharges). Finally,

we removed all regions that had 0 new hospitalization cases across all timepoints because we assumed that there was nothing to learn in these regions; here we set a low threshold because our search trends dataset only contained 12 regions after preprocessing and we wanted to keep as many regions as possible. The hospitalization dataset originally contained 56 regions in the United States and, after our processing, we were left with 42.

After merging the hospitalization dataset with the weekly search trends data set by region and by date, we were left with 10 regions and 121 symptoms.

The documentation for the dataset indicates that the relative popularity values were normalized with a region-specific value. Hence, in order to make comparisons between regions possible, we had to normalize our dataset. There are many ways to do this, and we initally considered re-normalizing the search trends data using a baseline value obtained from the weeks before COVID-19 started spreading in the United States; however, we decided that this would not be a very good baseline value because some search trends varied in accordance with seasonal changes. We tried dividing our data points by the mean of all data in the region it belonged to, but the PCA plots (not shown in this report) showed very distinct clusters (one for each region), and we thought that the data from each region was still too different. We ultimately decided to compute the percentage change from the mean (i.e., $\frac{value - mean}{mean}$, where *value* is the grand average of all data within a region, because we saw in the PCA plots that some regions were clustered together. We used the mean instead of the median because sometimes the median was 0, and we would get a division by 0.

We noticed that there existed a daily resolution version of the search trends dataset, and that it contained many more regions and symptom data than the weekly resolution one. Notably, it contained non-NaN data for the "cough" and "fever" symptoms (which had all missing values in the weekly resolution dataset), which are known to be COVID-19 symptoms. However, after converting this dataset to a weekly resolution and preprocessing it in a similar way as what we previously did (but we also removed features that had a very low variance), we found that, even though there was more data available in the daily dataset, it seemed noisier and would not necessarily be helpful for our purposes. We compared the features with the highest mean correlation with the hospitalization cases (averaged across regions) between the two datasets, and found that the top symptoms for the weekly resolution dataset – "ageusia" (loss of taste), "dysgeusia" (distortion of taste) and "anosmia" (loss of smell) – all seemed related to COVID-19, whereas the top symptoms for the daily resolution dataset included less COVID-19 related symptoms. Moreover, "cough" and "fever" were not present for the latter dataset.

3 Results

3.1 Data visualization and clustering

3.1.1 Search trend visualization

We used different methods when choosing the features to plot. We sorted our symptoms in four different methods and plotted heatmaps of the top 5 symptoms for each method. First, we computed the mean correlation (across all regions) between each symptom and the hospitalization cases in a region. We realized that, even though the search trends for these symptoms were highly correlated with the hospitalization cases for some regions, for others the values were constant and negative (i.e., before normalization they were all 0s). Because of this, we next looked at most correlated symptoms within each region, and chose those that occurred more frequently. For the last two methods we simply chose the symptoms that had the highest popularity (determined by summing up their normalized values over all regions) and highest variance (computed by pooling all regions together). We then plotted heatmaps for all of these symptoms (not shown in this report) and picked five of them that visually looked the most informative: "ageusia", "dysgeusia", "anosmia", "burning chest pain" and "shallow breathing". We plotted these alongside a plot showing normalized hospitalization numbers (percentage of the maximum in each region) for visual reference (Fig. 1A).

From these plots we can see that the hospitalization cases show similar patterns for some regions: namely, Rhode Island (US-RI), Maine (US-ME), New Hampshire (US-NH), Vermont (US-VT) and West Virginia (WV) all seem to show relatively high hospitalization cases in the first few months, while South Dakota (US-SD), North Dakota (US-ND), Wyoming (US-WY), Montana (US-MT) and Alaska (US-AK) show a very drastic increase in number of new hospitalizations later in the year (they also show a smaller increase in new hospitalizations around March). Interestingly, regions that were geographically close to each other showed similar trends.

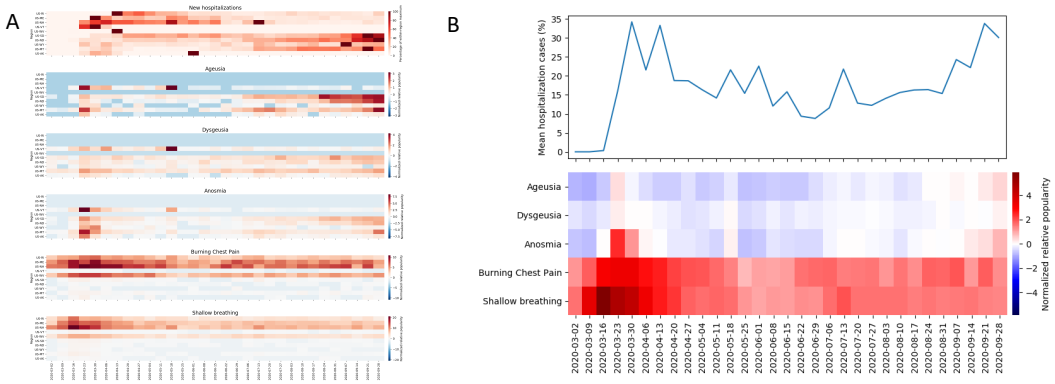


Figure 1: Task 2, symptom heatmaps

We also plotted the search trends for these symptoms and the hospitalization cases averaged across regions (Fig. 1B). The correlation between the symptom search trends and the peaks in the hospitalization cases (especially in March/April and in September) is visible here as well.

3.1.2 Dimensionality reduction with Principal Component Analysis

We visualized the dataset in a lower dimensional space using PCA. We found the first 8 components explain 90% of the variance (see Appendix). The Elbow Rule seemed to indicate that we should choose about 5 components (see Appendix), but we think the Elbow Rule is somewhat subjective and not very reliable, especially since the location of the "elbow" depends on the scale of our axes. Hence, we decided to use 8 PCs (based on cumulative variance ratio) in later steps.

We plotted the top 3 principal components (PCs) and used different colouring schemes (based on region, date and the logarithm of the number of hospitalizations)(Fig. 2A-C).

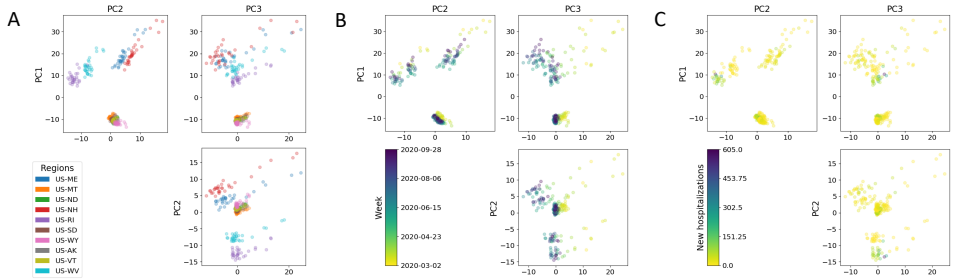


Figure 2: Task 2 principal components

We observed that PCA seemed to mostly group our data points by region, and that the points in some regions vary very similarly. Moreover, it seems like, along PCA, the data points are separated by their time points (positive values for earlier dates (in yellow), and negative values for later dates (in blue/purple).

3.1.3 Clustering

We used a K-Means model and created clusters for the original (normalized) dataset, for the PCA dataset (keeping all components) and for the PCA dataset with only the 8 first components. To determine the best number of clusters, we calculated the sum of squared errors relative to the centre of each cluster (see Appendix); we took the average from 10 iterations because different initializations produce different clusters. We also plotted the (average) silhouette score against the number of clusters (see Appendix). The silhouette score for a sample is computed as $\frac{b-a}{\max(a,b)}$, where a is the mean intra-cluster difference and b is the mean nearest-cluster distance; it can be used as a measure of how well defined our clusters are (6).

The sum-of-square errors plot was not very informative, as there is no clear point where it starts tapering off. The silhouette score plot, however, suggested that we use around 2-4 clusters.

We chose to use 3 clusters and projected our data onto the first two PCs in order to be able to visually compare the clusters (Fig. 4). We observed that the clustering is very conserved across the datasets. This is not very surprising because PCA is essentially just rotating the coordinate axes, and this doesn't change the distance between two points. When we choose a different k, sometimes there are small differences between the clustering in the 8-component PCA-reduced dataset, but the clustering is still very visually consistent.

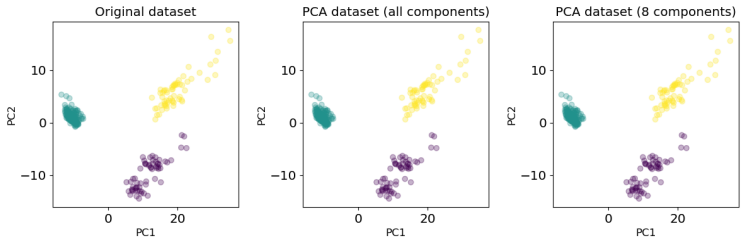


Figure 3: Task 2, clustering

3.2 Supervised learning

We performed supervised learning using different splitting methods for KNN and decision trees in order to predict the hospitalization cases given the search trends data. Results of KNN were determined by looking at the loss (computed as the mean squared error (MSE)) between the train set and the validation set at a varying number of neighbours. For decision trees, there were multiple parameters to vary. However, the results from varying maximum depth and leaves, as well as minimum features showed overfitting on training data and unstable MSEs on validation sets. Therefore, we decided to use minimal cost-complexity pruning from sklearn, which prunes the decision tree to avoid overfitting (5). Since every training set generated a different set of effective alpha values, the mean alpha values were used in order to keep the set of alpha values concise. Consequently, this compromised the results of some of the folds (see Appendix).

First, we evaluated the performance of both models by splitting the data by region. We performed 5-fold cross-validation on the data. We found that, for both KNN and decision tree models, not only was the cost high, but there was also a very high standard deviation in the validation sets while the training sets showed little to none (Fig. 4A). Upon investigating each fold separately, we found that the validation set that contains data for Rhode Island had the biggest MSE. This could be explained by the particularly high number of hospitalization cases in Rhode Island (the maximum was 605, while in the other regions no value exceeded 185) (see Appendix). Thus, overall high standard deviation in the validation set indicated that the model generated by splitting the data by region did not predict hospitalization cases very well and was unable to generalize what it learned to new regions.

Then, we evaluated the performance of KNN and decision tree models using testing sets based on date. Data points for dates before August 10, 2020 were put in the training set and the rest were in the test set. Compared to the KNN and decision trees trained on separate regions, both models performed more accurately (Fig. 4B). The mean squared error across any number of neighbours or value of alpha was generally smaller.

In addition, we repeated the analysis using subsets of the same dataset. We divided the dataset based on the K-means ($k = 3$) clusters from the previous task. We chose to use these clusters because they grouped similar regions together and each cluster contained at least two regions; we were particularly interested in seeing if this prediction strategy produced results that were better than what our region-based 5-fold cross-validation results. We used the same KNN and decision tree models to evaluate performance of the different splitting strategies. The performance of the decision tree model across the cluster varied a lot (see Appendix). This could be due to the small sample size (two of clusters only contained two regions each). Therefore, we could not determine the best performance based on this data.

4 Discussion and Conclusion

In this project, we applied unsupervised and supervised machine learning techniques on real-world datasets in order to predict cases of COVID-19 hospitalizations in the United States. We had to do extensive preprocessing of the

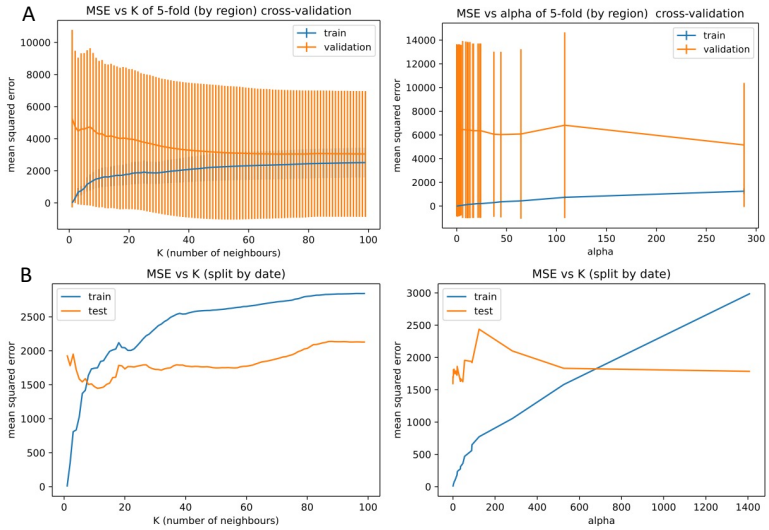


Figure 4. Performance of supervised learning.

a) The error produced from varying the number of K-neighbors in 5-fold cross-validation of KNN (left). The error produced from varying the value of alpha for pruning in 5-fold cross-validation of decision tree model (right).

b) The error produced from varying the number of K-neighbors of KNN by splitting the dataset by date (left). The error produced from varying the value of alpha for pruning of decision tree model by splitting the dataset by date (right).

Figure 4: Task 3, regression performance

data and make many decisions regarding the filtering criteria for symptoms and regions. We tried to make these decisions by first doing exploratory analyses and visualizations of the raw data (e.g., by plotting the distributions of specific measures when we wanted to set a threshold), but still found it generally difficult to find a good balance between having clean/informative data and having enough data for our subsequent analysis steps.

We were very interested by the fact that, in our preprocessed dataset, the symptoms whose search trends were the most correlated (on average) with COVID-19 hospitalization cases were indeed symptoms that have been reported in COVID-19 patients: loss of smell and/or taste, shallow breathing (7). We were surprised that, in the daily resolution dataset (which we preprocessed but did not use in later steps), cough and fever did not seem very correlated with hospitalization cases. One explanation we can think for this is that perhaps anosmia, ageusia and related symptoms are generally rarer (at least before COVID-19) than cough and fever (whose search trends show significant seasonal fluctuations in previous years as well), so maybe people tend to look up these symptoms more when they experience them.

PCA showed that our 121-feature dataset could be reduced to only 8 features while keeping 90% of the original variance. PCA seemed to group our data points mainly by region, and it also somewhat separated them by time (within each region) along one of the components (PC 3).

The K-means clusters were exactly the same between the original dataset and the PCA dataset with all components. When we used only the top principal components, the clusters were slightly different but still very consistent visually with the clusters in the 121-feature datasets.

Our supervised learning results were not very good, especially for the case where we did 5-fold cross-validation with the 10 regions (for both KNN and decision tree models). Since splitting the data by region gave a large error margin, the bad performance could be due to variability of search trends and hospitalization cases between regions. Therefore, the model trained on one region cannot predict the number of hospitalization cases of the search trend of another region. It would be interesting to compare performance within a sub-region of the United States. In fact, we tried to evaluate the performance of clusters, which grouped regions that are geographically close together, but due to small sample size, pruning was inefficient. On the other hand, the loss in decision trees in the larger dataset were overall much smaller than in KNN. This means that search trends are less dependent on time. The smaller errors in predicting cases by date means that the model can be be trained and predict from up-to-date data.

5 Statements of contribution

Anny wrote the code for Task 3 and contributed to the Introduction and Results sections of the write-up. Sameen contributed in Task 1 and in the Datasets and Results sections of the write-up. Michelle wrote the code for Task 1 and Task 2 and contributed to the Datasets, Results and Discussion sections of the write-up.

References

1. Nuti, Sudhakar V et al. "The use of google trends in health care research: a systematic review." PloS one vol. 9,10 e109583. 22 Oct. 2014, doi:10.1371/journal.pone.0109583
2. Teng, Yue et al. "Dynamic Forecasting of Zika Epidemics Using Google Trends." PloS one vol. 12,1 e0165085. 6 Jan. 2017, doi:10.1371/journal.pone.0165085
3. Everett, Katie, et al. "Open COVID-19 Data." Github, github.com/google-research/open-covid-19-data/blob/master/data/exports/search_trends_symptoms_dataset/United%20States%20of%20America/2020_US_weekly_symptoms_dataset.csv.
4. Everett, Katie, et al. "Google-Research/Open-Covid-19-Data Katie Everett, Dan Nanas, Maddy Myers (UCSD), Sumit Arora, and Ian Fischer." GitHub, 19 Oct. 2020, github.com/google-research/open-covid-19-data blob/master/data/exports/cc_by/aggregated_cc_by.csv. "1.10. Decision Trees." Scikit, scikit-learn.org/stable/modules/tree.html.
5. "Post pruning decision trees with cost complexity pruning." Scikit-learn. https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html
6. "Selecting the number of clusters with silhouette analysis on KMeans clustering." Scikit-learn. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
7. Wagner T, et al. "Augmented Curation of Clinical Notes from a Massive Ehr System Reveals Symptoms of Impending Covid-19 Diagnosis." Elife, vol. 9, 2020, doi:10.7554/eLife.58227.

Appendix

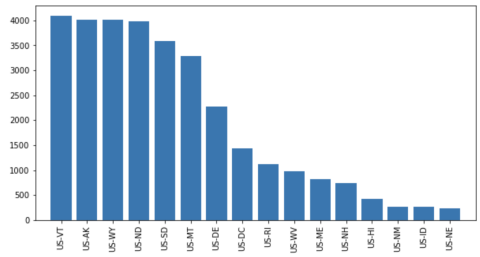


Figure: Regions with 500 or more non-NaN rows

symptom:Ageusia	0.482122
symptom:Dysgeusia	0.373866
symptom:Anosmia	0.368366
symptom:Ascites	0.324004
symptom:Eye pain	0.237695
symptom:Hepatic encephalopathy	0.175625
symptom:Depersonalization	0.169549
symptom:Hypercalcaemia	0.164180
symptom:Hydrocephalus	0.160441

Figure: Correlation between regions and new hospitalization cases (weekly resolution dataset)

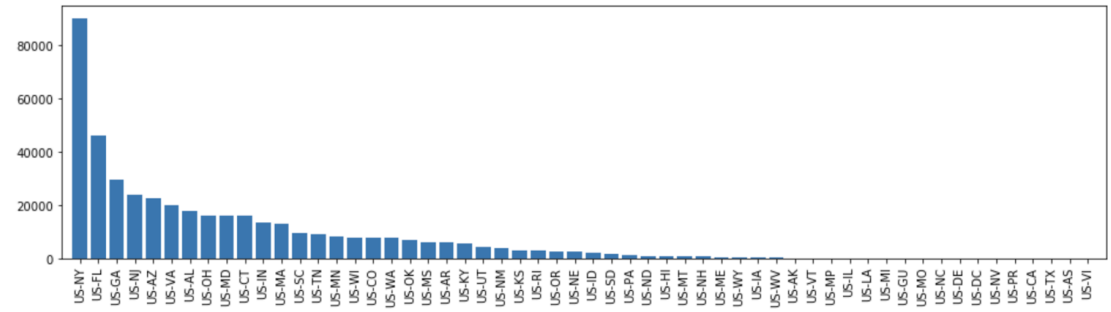


Figure: Regions and corresponding hospitalization cases

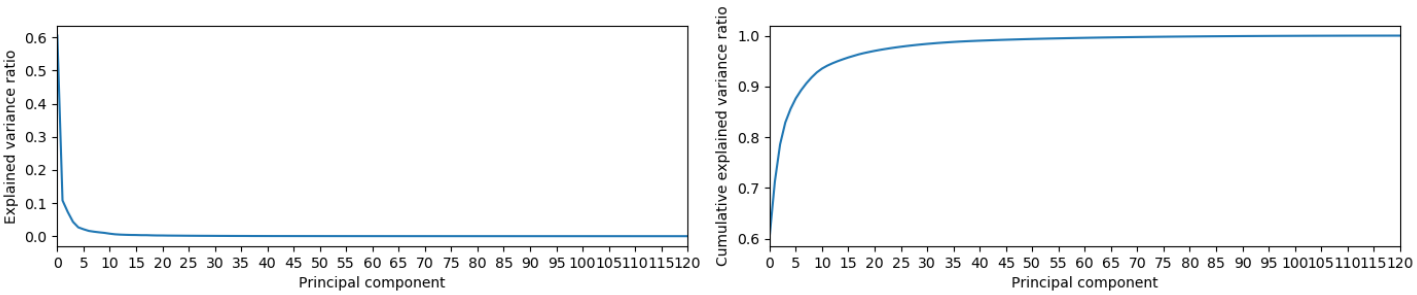


Figure: Explained variance ratio and cumulative variance ratio vs principal component

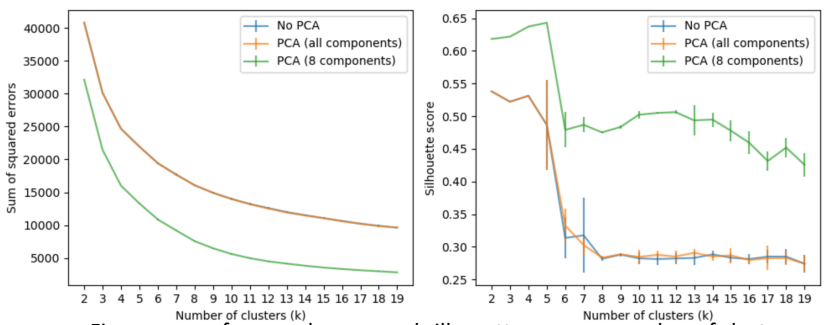


Figure: sum of squared errors and silhouette score vs number of clusters

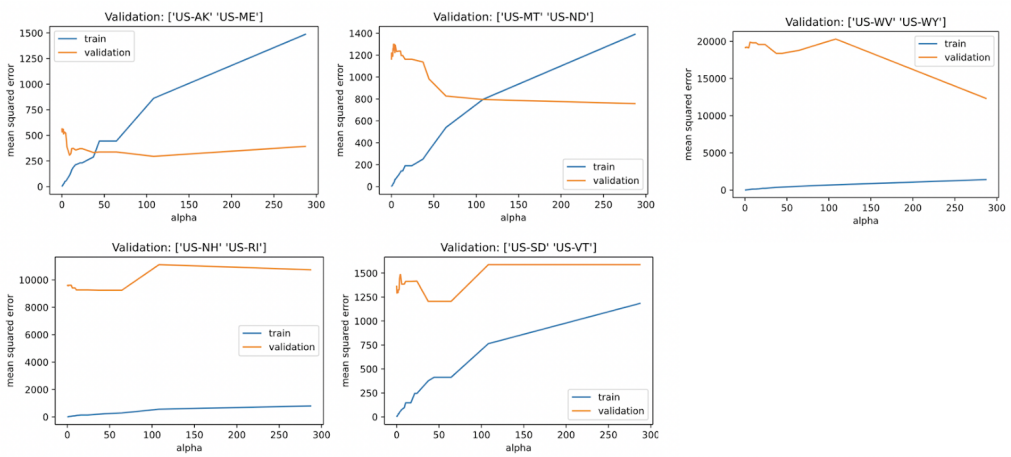
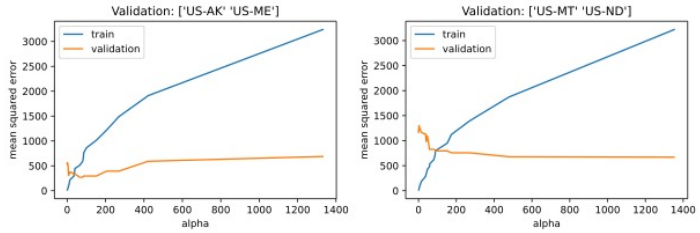
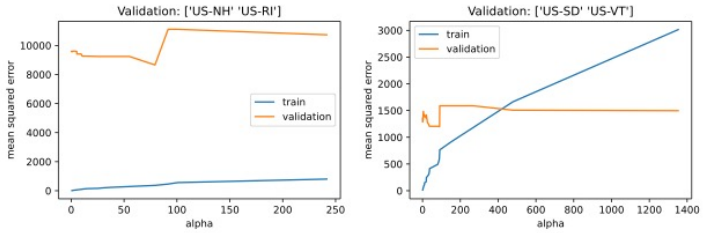
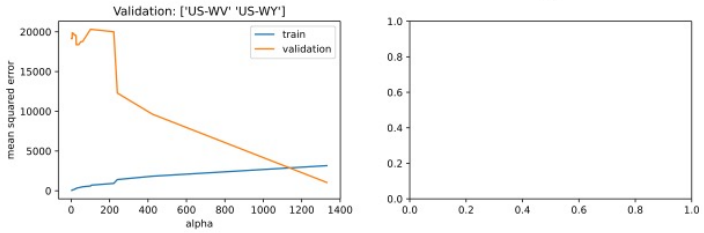
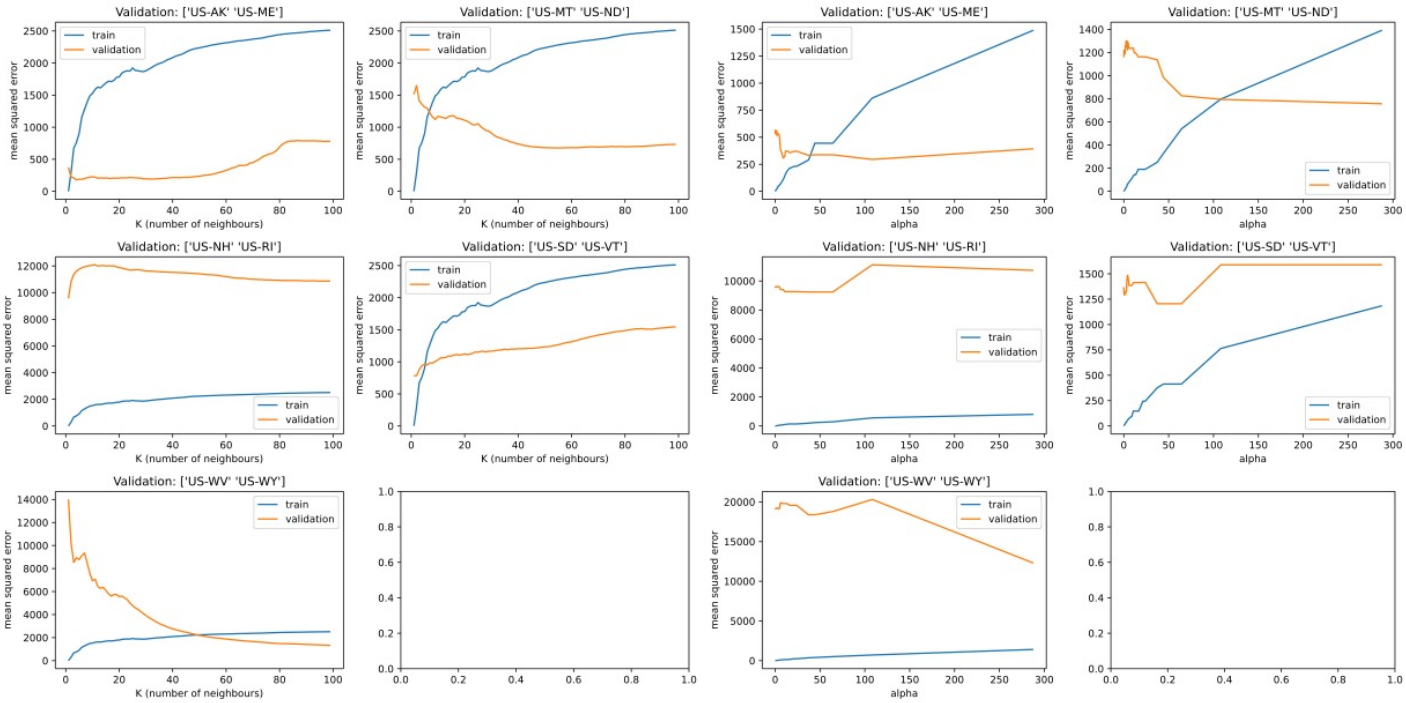


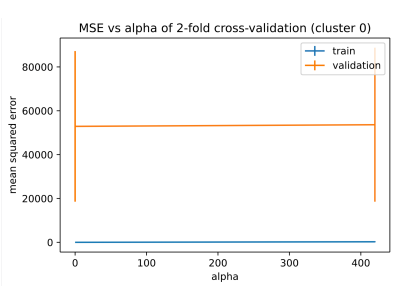
Figure: Mean squared ratio vs alpha for each region

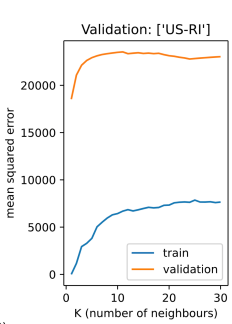

Validation: ['US-AK' 'US-ME']
Validation: ['US-MT' 'US-ND']
train
validation
mean squared error
alpha

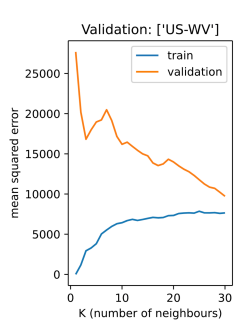

Validation: ['US-NH' 'US-RI']
Validation: ['US-SD' 'US-VT']
train
validation
mean squared error
alpha

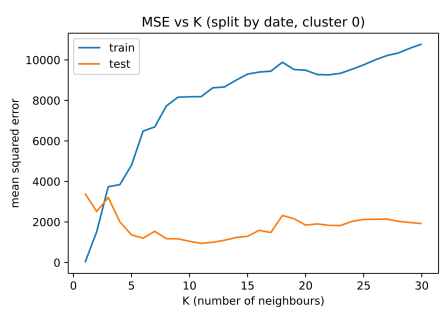

Validation: ['US-WV' 'US-WY']
train
validation
mean squared error
alpha

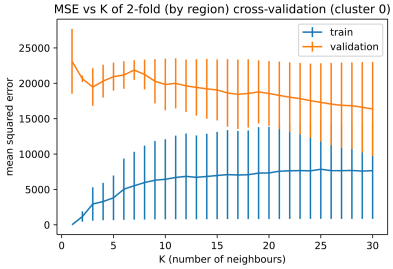

Validation: ['US-AK' 'US-ME']
Validation: ['US-MT' 'US-ND']
Validation: ['US-NH' 'US-RI']
Validation: ['US-SD' 'US-VT']
Validation: ['US-WV' 'US-WY']
train
validation
mean squared error
K (number of neighbours)
alpha

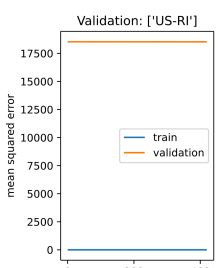

MSE vs alpha of 2-fold cross-validation (cluster 0)
train
validation
mean squared error
alpha

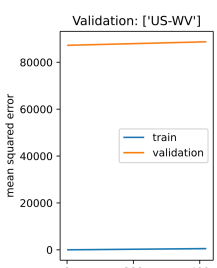

Validation: ['US-RI']
train
validation
mean squared error
K (number of neighbours)

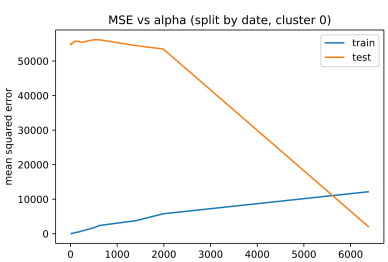

Validation: ['US-WV']
train
validation
mean squared error
K (number of neighbours)


MSE vs K (split by date, cluster 0)
train
test
mean squared error
K (number of neighbours)


MSE vs K of 2-fold (by region) cross-validation (cluster 0)
train
validation
mean squared error
K (number of neighbours)


Validation: ['US-RI']
train
validation
mean squared error
alpha


Validation: ['US-WV']
train
validation
mean squared error
alpha


MSE vs alpha (split by date, cluster 0)
train
test
mean squared error
alpha