Assignment 2 - Report

STAT497

Presented to

Dr. Frederic Godin

By

Nam Phuong Nguyen (40011040)

Tony Yuan (40029336)

Spyros Orfanos (40032280)

March 22, 2018

Question 1

**Part (a)**

Estimating Coefficients:

The probability of a default follows a Bernoulli distribution with probability *p1(x).* For the logistic regression model, we have *p1(x) =*P(Y=1|X=*x*) = , where η*i* = .

We estimate the logistic regression coefficients using Maximum Likelihood Estimation:

\*

This is equivalent to minimizing the negative log-likelihood, which simplifies to:

)

Coefficients:

-5.675913 (Intercept)

-1.713641 (MaritalStatus)

2.253775 (LTV)

23.316316 (DiffRate)

Interpretation of Coefficients:

1. Marital Status

* The coefficient is negative meaning that if someone is married they are less likely to default. This is consistent with the intuition: a married couple usually has a higher household income than a single person, so the loan can be repaid more easily.

1. LTV

* The coefficient is positive meaning the higher the LTV the higher the more likely a mortgage default. This is consistent with the intuition: the higher the LTV, the more money was lent for the mortgage. It is more difficult to repay a large amount of money than a small amount, so when LTV increases the person is less likely to repay a loan.

1. DiffRate

* The coefficient is positive so the higher the DiffRate the more likely a mortgage default. This is consistent with the intuition: a large DiffRate means that the interest rate has increased, and so the interest payments will be larger than initially expected. Thus, it would be more difficult to repay a loan, resulting in a higher probability of default.



**Part (b)**

The error rate is small: only 2.27% of the observations are misclassified. However, this does not necessarily mean that the classifier is a good one. Firstly, the error rate was calculated on the training data, so the test error rate could be significantly higher. Also, only 2.26% of the individuals in the sample defaulted on their mortgage, so even the null classifier which always predicts that an individual will not default results in the same error rate as our model. Of the individuals who defaulted, none of them were classified as default, so there is a 100% error rate for this group of individuals. So, this classifier is not a good one.

**Part (c)**

A = 30

Confusion Matrix:

Error Rate: 29.77%

A = 60

Confusion Matrix:

Error Rate: 46.13%

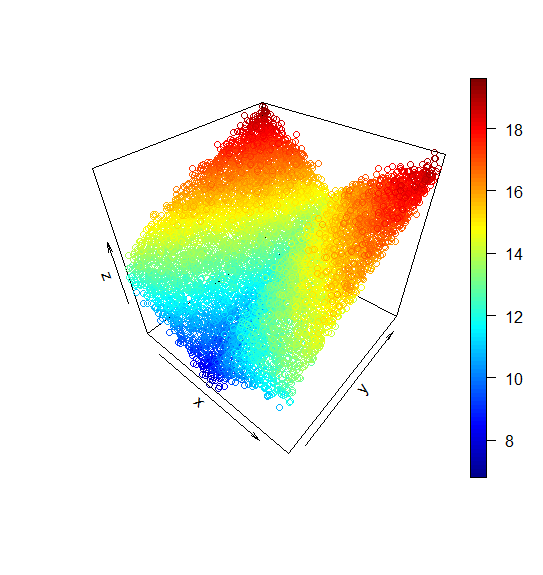
As A increases, we observe that the number of false negatives decreases as this type of misclassification has a greater penalty.

**Part (d)**



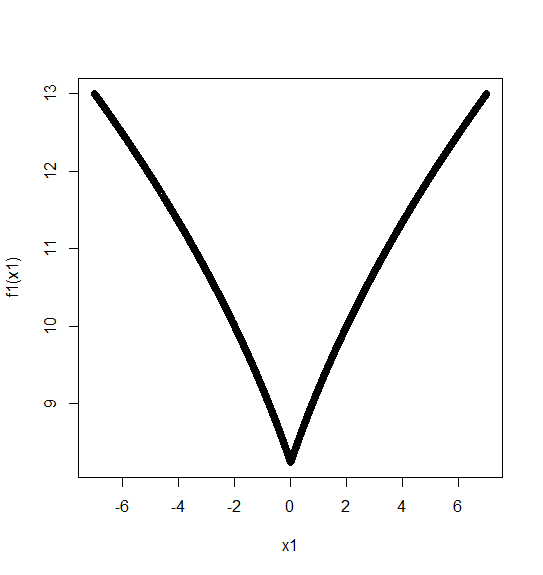
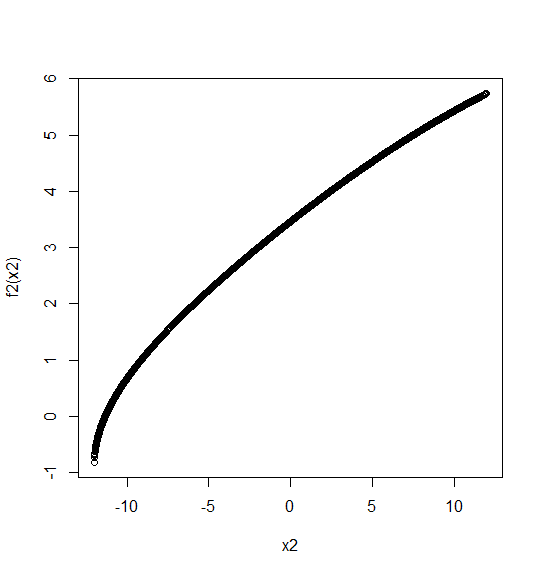
* The area under the ROC curve for this model is **0.7632363**
* The predictive power of the model is rated as “Acceptable”
* Note: the area was approximated using the trapezoid method and 5000 values of A
* A ROC curve for a logistic regression model with no predictive power would increase to 1 at a slow rate since there would be many misclassified observations resulting in many False Positives and False Negatives. Having many false positives results in a lower TPR, and having many false negatives results in a high FPR. As A increases, the model will classify more observations as Positive, so FPR and TPR increase and converge to 1.
* A ROC curve for a logistic regression model with very strong predictive power would initially increase at a very fast rate since the model would accurately classify observations, resulting in a high rate of true positives (and a high rate of true negatives). As A increases, the model will classify more observations as Positive, so FPR and TPR increase and converge to 1.

Question 2

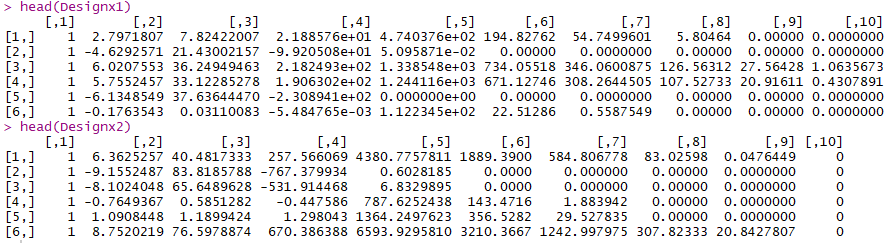
**Part (a)**

I generated the two functions such that they both can be used on a vector and return the appropriate vector with the new values that have been passed through the function. The scatter3d plot of the x and y values define the relationship between this function.

For the two graphs below we have graphed the relationship between x1 and f1(x1) and x2 and f2(x2) below. We will try to replicate these with the curves.

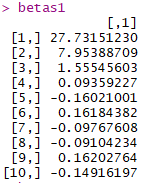
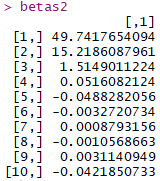


**Part (b)**

The function for creating the design matrix makes sure that all h(x,ξ) = (x - ξ )^3 becomes 0 if it is negative. Each value of ξ is one of the nodes that were created in the sequence for this problem. 

**Part (c)**

The beta values were calculated using a recursive method through back fitting. Upon the 100th iteration these values of beta were generated for beta 1 and 2.

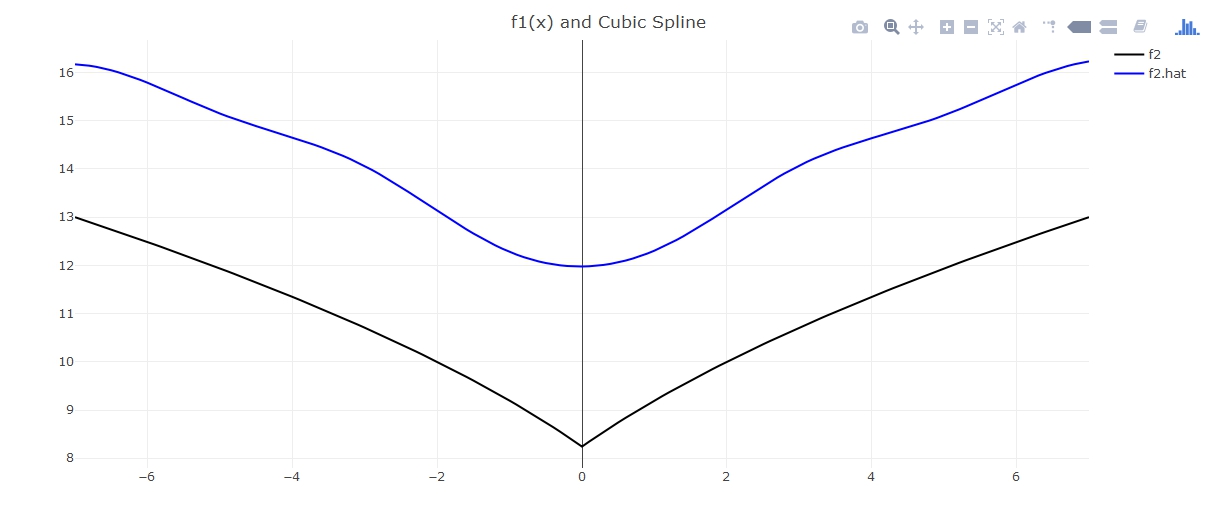
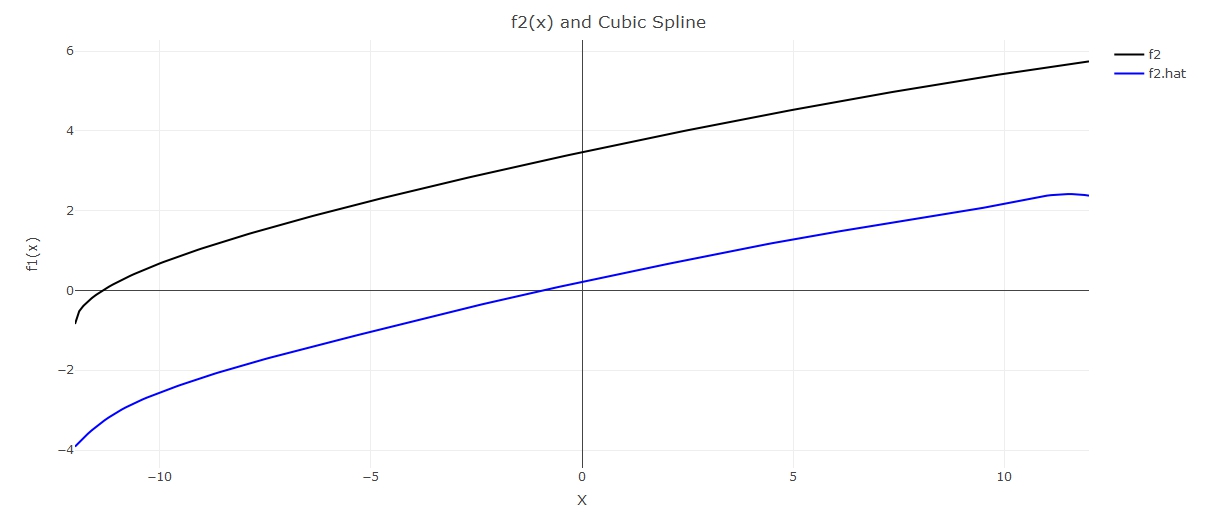
 

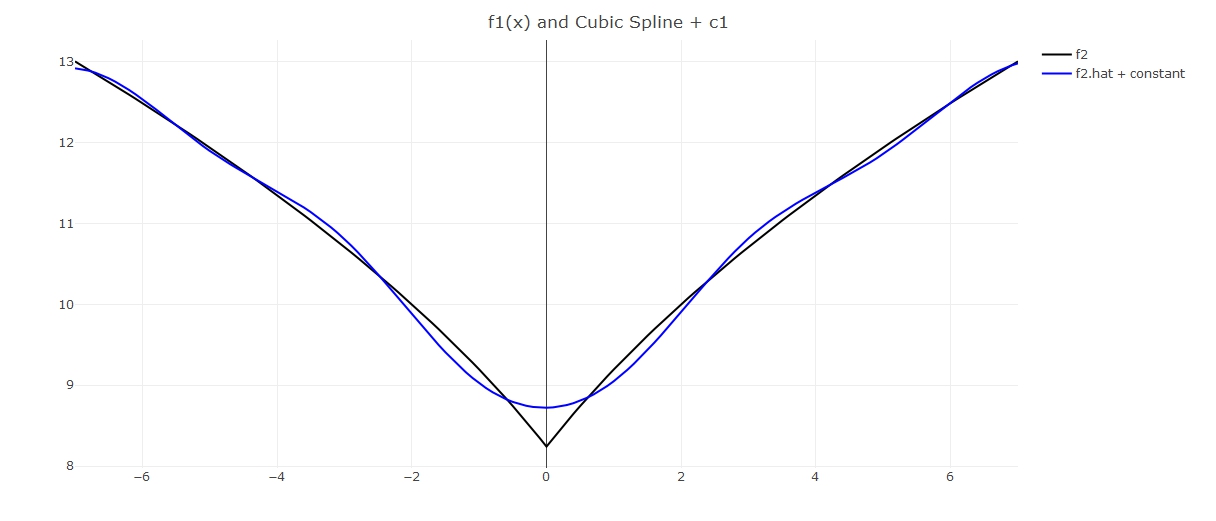
**Part (d)**

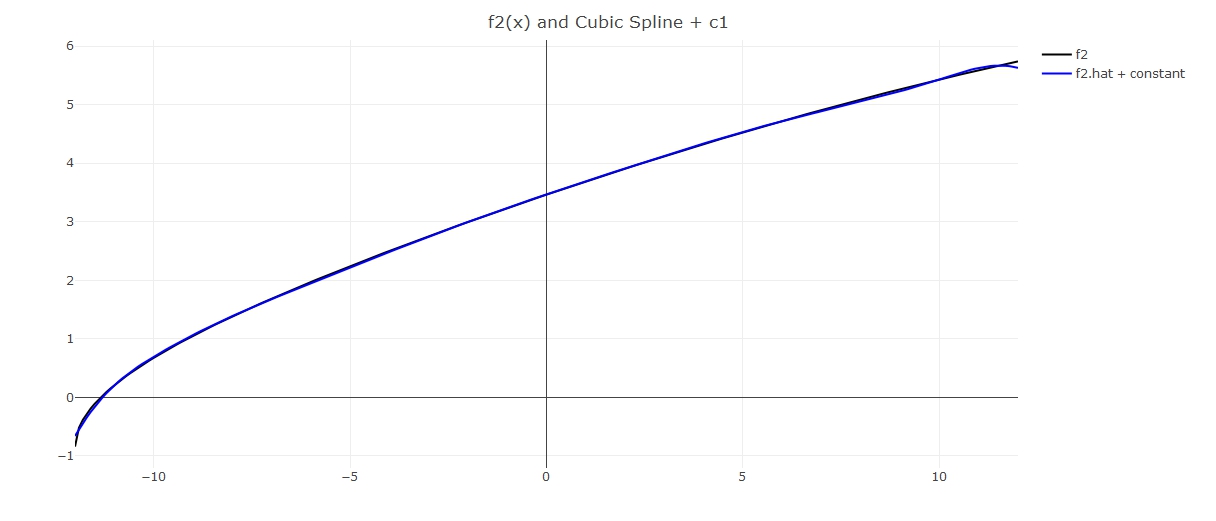
Graphing and comparing the curves of f1 and f2 with the cubic splines we can see that the cubic splines are very similar in terms of the shape of the graph but that there is a translation between the actual curve and the cubic spline. The reason that the function is not the same is the actual function has an exact formula using absolute values and square roots while our formula is a polynomial so thus our polynomial is doing its best ability to approximate our function which is why they aren’t the same. It is offset because when using the backfitting algorithm there is an alpha that needs to be calculated that is a constant to be added.



Once we calculate c1 and use it to translate our cubic splines we can see that the actual curve and the cubic spline are very similar.







Question 3

**Part (a)** - See code

**Part (b)**

Kmeansclustering function which takes in a nxp *Datapoints* matrix, a *ncluster* (call it k) integer representing the number of clusters we want, a *theseed* integer and *maxiter* integer which is the maximum number of iterations we want he algorithm to perform.

|  |  |
| --- | --- |
| Step 0: | **Set the seed** |
| Step 1: | **Randomly generate the centroids.** For each of the p features (each column of *Datapoints*), we generate a pseudouniform random number between the minimum and the maximum of that feature across all observations using the runif function. This results in a k by p matrix called *centroid.* |
| Step 2: | **Assign a cluster to each of the n observation.** For each observation, we calculate its Euclidean distance with each of the clusters and store it in the k length vector *dist*. Determine using the function *which* which of the k distances is the smallest and assign that cluster to the observation. The cluster assignments are stored in the *cluster1* vector of length n. |
| Step 3: | **Iterations to determine the clusters.** These steps are done a maximum of *maxiter* times.   1. For each cluster, we make a temporary matrix which contains only the observations assigned to that cluster. Recalculate the centroids by taking the mean of each feature across the observations in the cluster. If a cluster is empty, we keep the same coordinates. 2. Repeat Step 2 to assign a new cluster (stored in *cluster2)* to each observation. 3. If *cluster1* and *cluster2* are identical, end the iteration. Else, migrate the values in *cluster2* to the *cluster1* vector. |
| Step 4: | **Return KmeansOut**, a list containing the final cluster assignments in a length n vector and the centroids, a k by p matrix. |

**Part (c)**

Since the final centroids and clusters depend on seed used to randomize the initial centroids, we perform the algorithm until all clusters have at least one observation using different seeds for the number of clusters we want K=2,3,4.

**ANALYSIS OF RESULTS**

From Table 1.2, we can see that only COMED is in a different zone, meaning that it has a different electricity price as the other zones. When looking at the map, we realize that COMED contains Chicago which is the biggest city in the region and probably consumes the most electricity compared to the other cities. Hence, the price would be more expensive.

From Table 2.2, we can see that again COMED is placed in its own cluster for the same reason as in Table 1.2. However, AEP and EKPC are placed in their own cluster meaning that they have similar electricity prices. On the map, they are seen situated right next to each other.

From Table 3.2, we can see that COMED, again, stands on its own. MID-ATL/APS, DPL, DEOK, PEPCO and BGE are part of the same cluster. These are however all over the place in the region which means that it doesn’t seem that their geographical placement has anything to do with the price. AECO, AEP, DOM, EKPC, JCPL, PECO, PPL, PSEG, RECO are in the same cluster. APS, ATSI, DAY, DUQ, METED, PENELEC are part of another cluster. Again, they aren’t necessarily close to each other.

Across the different K’s, we can see that COMED is always put in its own cluster, meaning that its electricity price is significantly different than the other zones. Nothing else of subsistence can be said about the other zones concerning how different their electricity pricing is.

**SUMMARY OF RESULTS**

Table 1.1 : Final centroids for K=2

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] | [,11] | [,12] |
| [1,] | 26.92889 | 26.35788 | 25.94618 | 26.53489 | 27.814 | 31.80206 | 50.42631 | 50.64016 | 49.30781 | 42.52653 | 43.38721 | 43.99883 |
| [2,] | 23.15534 | 22.43232 | 22.19847 | 22.70389 | 23.79844 | 25.97244 | 38.09921 | 40.40585 | 37.43916 | 31.44977 | 32.0529 | 32.64547 |
|  | [,13] | [,14] | [,15] | [,16] | [,17] | [,18] | [,19] | [,20] | [,21] | [,22] | [,23] | [,24] |
| [1,] | 41.22501 | 39.28858 | 37.75263 | 37.50096 | 42.5009 | 54.7057 | 60.11553 | 54.49591 | 50.93594 | 47.22961 | 44.1754 | 39.77018 |
| [2,] | 29.67697 | 27.89946 | 25.79995 | 24.87202 | 27.48519 | 29.61236 | 31.17237 | 29.2593 | 28.04866 | 26.41958 | 23.44521 | 17.4372 |

Table 1.2 : Final cluster assignments for each zone for K=2

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MID-ATL/APS | AECO | AEP | APS | ATSI | BGE | COMED | DAY | DEOK | DOM |  |
| 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |  |
| DPL | DUQ | EKPC | JCPL | METED | PECO | PENELEC | PEPCO | PPL | PSEG | RECO |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Table 2.1 : Final centroids for K=3

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] | [,11] | [,12] |
| [1,] | 27.17199 | 26.61671 | 26.18141 | 26.75577 | 28.0146 | 32.02566 | 50.73382 | 50.85107 | 49.53202 | 42.70993 | 43.65048 | 44.34464 |
| [2,] | 23.15534 | 22.43232 | 22.19847 | 22.70389 | 23.79844 | 25.97244 | 38.09921 | 40.40585 | 37.43916 | 31.44977 | 32.0529 | 32.64547 |
| [3,] | 24.74102 | 24.0284 | 23.82917 | 24.54706 | 26.00852 | 29.78961 | 47.65869 | 48.74202 | 47.28998 | 40.87591 | 41.01778 | 40.88652 |
|  | [,13] | [,14] | [,15] | [,16] | [,17] | [,18] | [,19] | [,20] | [,21] | [,22] | [,23] | [,24] |
| [1,] | 41.65165 | 39.73751 | 38.19763 | 37.92126 | 43.01049 | 55.57215 | 60.75575 | 54.91318 | 51.2249 | 47.57269 | 44.52528 | 40.38111 |
| [2,] | 29.67697 | 27.89946 | 25.79995 | 24.87202 | 27.48519 | 29.61236 | 31.17237 | 29.2593 | 28.04866 | 26.41958 | 23.44521 | 17.4372 |
| [3,] | 37.38525 | 35.24824 | 33.74764 | 33.71827 | 37.91457 | 46.90766 | 54.35357 | 50.74052 | 48.33532 | 44.14193 | 41.02645 | 34.27182 |

Table 2.2: Final cluster assignments for each zone for K=3

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MID-ATL/APS | AECO | AEP | APS | ATSI | BGE | COMED | DAY | DEOK | DOM |  |
| 1 | 1 | 3 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |  |
| DPL | DUQ | EKPC | JCPL | METED | PECO | PENELEC | PEPCO | PPL | PSEG | RECO |
| 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Table 3.1: Final centroids for K=4

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] | [,11] | [,12] |
| [1,] | 26.82455 | 26.23366 | 25.78059 | 26.37129 | 27.62758 | 31.43832 | 48.68718 | 48.96456 | 47.35087 | 40.92526 | 42.22064 | 43.0583 |
| [2,] | 26.56975 | 26.02423 | 25.67024 | 26.27964 | 27.55022 | 31.67786 | 50.73398 | 51.85704 | 50.59564 | 43.62676 | 43.80061 | 44.21537 |
| [3,] | 27.54767 | 26.98185 | 26.57538 | 27.13568 | 28.46608 | 32.60583 | 53.18754 | 52.19601 | 51.28492 | 44.08852 | 44.99096 | 45.43193 |
| [4,] | 23.15534 | 22.43232 | 22.19847 | 22.70389 | 23.79844 | 25.97244 | 38.09921 | 40.40585 | 37.43916 | 31.44977 | 32.0529 | 32.64547 |
|  | [,13] | [,14] | [,15] | [,16] | [,17] | [,18] | [,19] | [,20] | [,21] | [,22] | [,23] | [,24] |
| [1,] | 40.54209 | 38.75403 | 37.25905 | 37.00295 | 41.7914 | 53.06006 | 57.2488 | 52.04661 | 48.77945 | 45.52873 | 42.85904 | 39.1376 |
| [2,] | 41.09628 | 39.04644 | 37.46757 | 37.20766 | 42.23321 | 53.77346 | 59.59083 | 54.67021 | 51.64014 | 47.91265 | 44.2849 | 38.09271 |
| [3,] | 42.60874 | 40.54135 | 38.98314 | 38.74934 | 44.09923 | 58.78655 | 65.90527 | 58.6955 | 53.97259 | 49.47155 | 46.41344 | 42.92179 |
| [4,] | 29.67697 | 27.89946 | 25.79995 | 24.87202 | 27.48519 | 29.61236 | 31.17237 | 29.2593 | 28.04866 | 26.41958 | 23.44521 | 17.4372 |

Table 3.2: Final cluster assignments for each zone for K=4

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MID-ATL/APS | AECO | AEP | APS | ATSI | BGE | COMED | DAY | DEOK | DOM |  |
| 3 | 1 | 1 | 2 | 2 | 3 | 4 | 2 | 3 | 1 |  |
| DPL | DUQ | EKPC | JCPL | METED | PECO | PENELEC | PEPCO | PPL | PSEG | RECO |
| 3 | 2 | 1 | 1 | 2 | 1 | 2 | 3 | 1 | 1 | 1 |