Assessing the aging infrastructure through data-mining of the National Bridge Inventory: an exploratory analysis

Alejandro Belenguer

2018-12-22

Contents

Chapter 1

Introduction

The aging trend of the U.S. bridge inventory has been arisen by the ASCE in the recent years (ASCE, 2017). An increasing population of older bridges creates a challenging scenario for the future bridge maintenance strategy. However, the observed reduction of the structurally defficient bridges is a first step forward.

Figure 1.1 shows the formentioned aging effect. The data has been downloaded from the publicly accessible National Bridge Inventory (NBI) of the U.S. Federal Highway Administration. The horizontal shift between peak densities represent the time lag (26 years), while the vertical shift shows a decrease in the total population density.

Stablishing the right criteria to sustain bridge condition on safe levels is a key factor in a problem of limited resource allocation. Thus, the study of the relation between bridge characteristics and bridge performance can uncover better bridge maintenance policies.

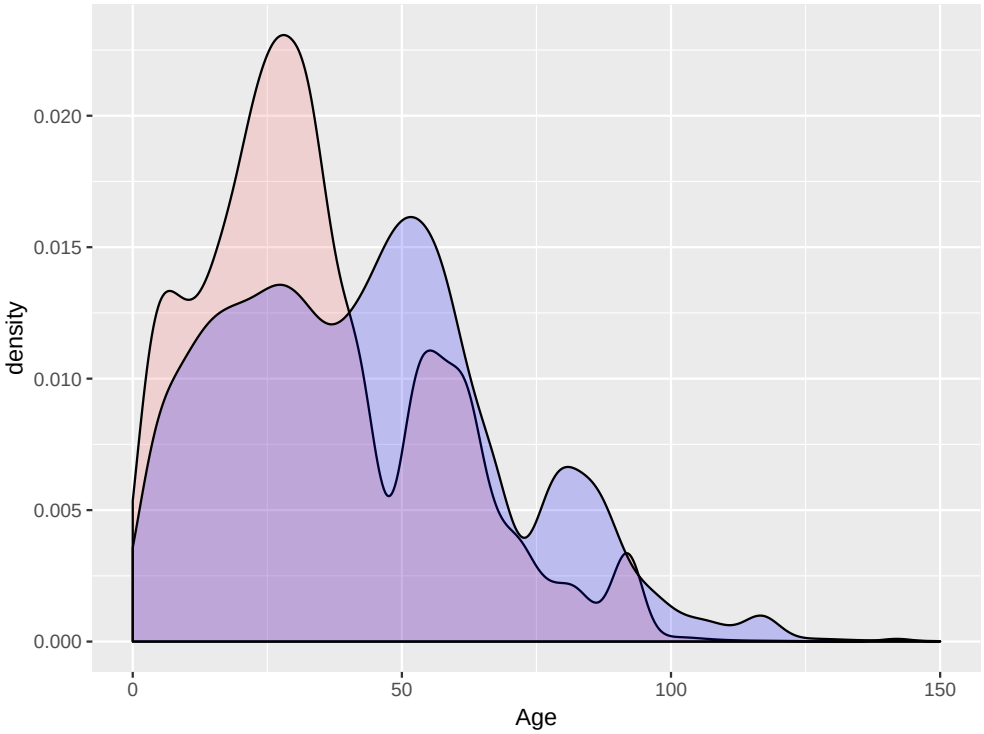

Figure 1.1: Age distribution of US bridges (1992 vs. 2017)

Chapter 2

Previous techniques and methods

The NBI has been previously exploited to predict bridge ratings, i.e. structural condition, from different elements such as deck, superstructure, substructure, or culvert. The vast amount of data available, totalling more than 600,000 bridges, has been tackled by subsetting the inventory into smaller samples using different criteria. Contreras-Nieto (2014) used the steel and prestressed concrete bridges in Oklahoma, while Saeed et al. (2017) worked on the Indiana database.

Related studies have been carried out using neural networks, Markov models, and regression trees (Veshosky et al., 1994) to the bridge superstructure condition. Bektas Basak Aldemir et al. (2013) applied classification and regression trees to predict bridge ratings and later on used recursive partitioning Bektaş (2017).

According to Contreras-Nieto et al. (2018), in the majority of the previous studies the model validation was not performed and the results showed a low adjusted R^2 value (around 0.4). This problem resulted in the unability of the models to predict structurally deficient ratings, as they only represent a small fraction (6% - 9%) of the inventory.

Chapter 3

Research problem

This work tries to provide a better insight into the topic by using novel data mining techniques. In particular, three non-supervised learning techniques will be used to generate "statistically similar" subsets of data.

The first one used is the clustering of extremes (partition around medioids) technique (Bracken et al., 2015). As a second approach, a classical Principal Component Analysis is carried out. Finally, Self-Organizing Maps will be used to compare with previous results.

With the results from the non-supervised approach, a multinomial regression will be used to predict the strucural condition of the bridges, and compare those with regression models without previous clustering.

3.1 Description of the data

The NBI database accounts for more than 136 parameters of the bridge inventory gathered at each bridge inspection. More information about the methodology can be consulted in the Recording and coding guide for the structure inventory and appraisal of the nation's bridges Weseman (1995).

The forementioned previous work found a fraction of those variables to be statistically significant when using regression models to predict bridge condition. Working from those, the author has selected the same to guarantee enough breadth of scope in the analysis.

The table 3.1 summarizes the variables names and numeric type adopted departing from Weseman (1995). In order to simplify the anlysis, some variables have been transformed, according to the following description:

1. Latitude, Longitude: considered full numeric precision available (hundredths of a second).
2. Year built: used to calculate bridge age.
3. ADT: considered full precision. However, the variable considered has been the Truck Average Daily Traffic, as it is the one considered significant (Saeed et al., 2017). The TADT is obtained multiplicating the ADT by the percentage of trucks in ADT (code 109).
4. Design load. Transformed to binary. 1 if code is known, 0 if not.
5. Structure kind and type: the original data considers different building materials for the first case, and different bearing mechanisms for the latter. The bridge selection process lead to consider only three kinds of superstructure material: steel, reinforced concrete, and prestressed concrete. Similarly, the most frequent structural type was girder / multibeam bridge. Thus, only this category was retained.
6. Length of maximum span: only bridges between 15 and 50 m. of maximum span length have been considered with the objective of comparing similar structures. Within the range, full numeric values are used. Figures 3.1 and 3.2 show the process followed.

Table 3.1: Selected variables

Variable	Code number	Numeric type	Temporal type
Structure number	008	Character	Stationary
Latitude	016	Numeric	Stationary
Longitude	017	Numeric	Stationary
Year built	027	Numeric	Stationary
Average Daily Traffic	029	Numeric	Time series
Design load	031	Binary	Stationary
Service under bridge	042B	Categorical	Stationary
Structure kind	043A	Categorical	Stationary
Structure type	043B	Categorical	Stationary
Length of maximum span	048	Numeric	Stationary
Structure length	049	Numeric	Stationary
Bridge roadway width	051	Numeric	Stationary
Deck Condition	058	Categorical	Time series
Superstructure condition	059	Categorical	Time series
Substructure condition	060	Categorical	Time series
Culvert condition	062	Categorical	Time series
Year reconstructed	106	Binary	Time series
Percentage of Truck in ADT	109	Numeric	Time series

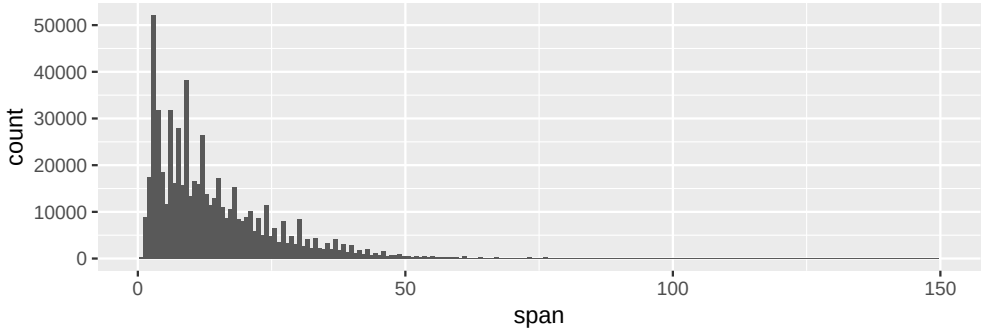

Figure 3.1: Max. span length distribution of selected bridges

7. Structure length: The wide range of lengths was transformed into four bins: from 15 to 50 m., from 50 to 100 m., from 100 to 200 m., and longer. The median length values were used to replace previous values (25, 75, 150, and 500 m., respectively).

8. Bridge total width: full numeric precision considered.

9. Deck, superstructure, substructure, and culvert condition: Original data adopted a 0 to 9 scale to classify the specific bridge condition. However, a new variable considering the structural deficiency has been used to reflect the definition given by the Federal Highway Administration. The term is defined as the classification given to a bridge which has any component (Item 58, 59, 60, or 62) in Poor or worse condition (code of 4 or less).

10. Year reconstructed: A fraction of the bridges has had major interventions that exceed the regular maintenance practice. A binary variable indicating if a reconstruction has existed at a given year has been used to increase regression accuracy.

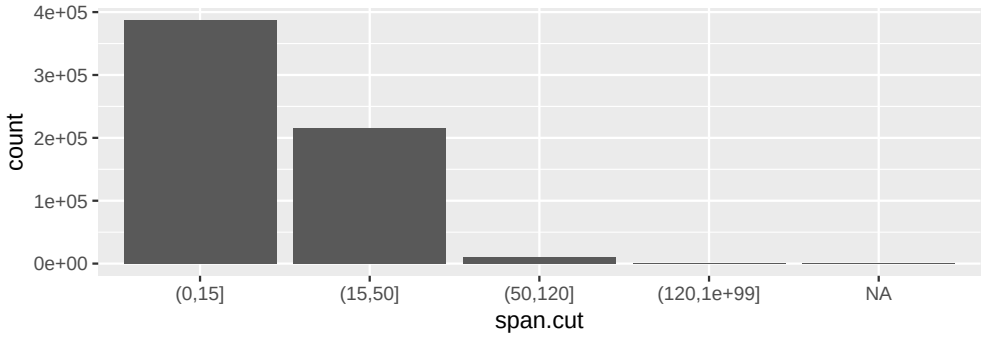

Figure 3.2: Max. span length bins used to select bridges for the analysis

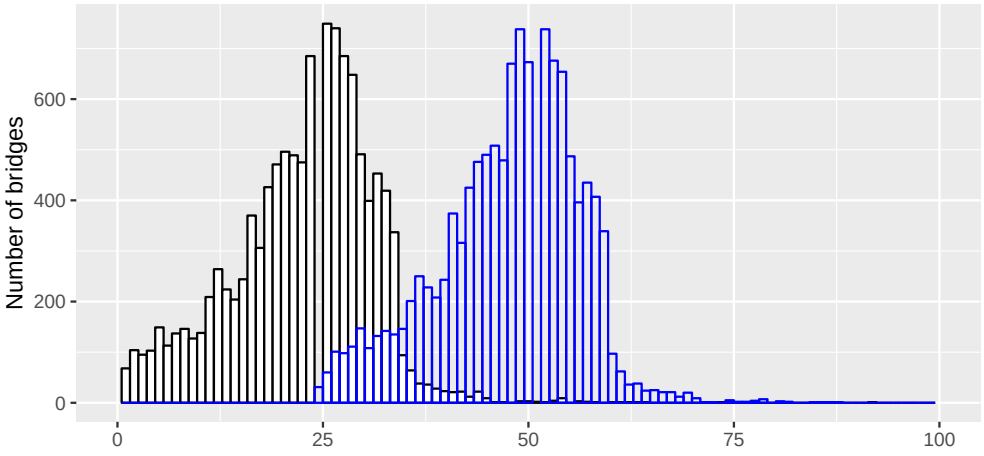

Figure 3.3: Age of selected bridges (1992 vs. 2017)

3.2 Hypothesis and diagnostics

The selected covariates were used to generate a subset of the entire dataset. The criteria followed to reduce the number of case studies was directly related to quality control and easiness of manipulation.

First, only bridges with the same name in 1992 and 2017 were retained. All bridges been replaced or renamed were consequently excluded. Additionally, only bridges with known location were included. Note that older bridges are a consequence of choosing this criteria, as shown in figure 3.3.

Second, the condition rating of the selected bridges had to be known. The values were in a few cases omitted for certain intermediate years. In this cases the previous known value was used to generate continuous data. Figures 3.4 and 3.5 depict the how fewer structurally deficient bridges and greater close-to-deficiency bridges phenomena occur simultaneously.

Third, the maximum span length, structure kind, and structure type matched the criteria described above. Only 15-50 max. span, steel/concrete/prestressed concrete, multibeam bridges were then considered. Figures 3.6, 3.7, and 3.8 show the property distribution on the selected bridges.

Finally, only bridges in the continental US STRANHET corridors were used. The STRAHNET corridor is formed by those highways considered strategically important to the defense of the United States.

A total sample of 12,970 bridges scattered throughout the entire continental US was used in the analysis, including 26 years of record. Figure 3.9 maps them on the US territory.

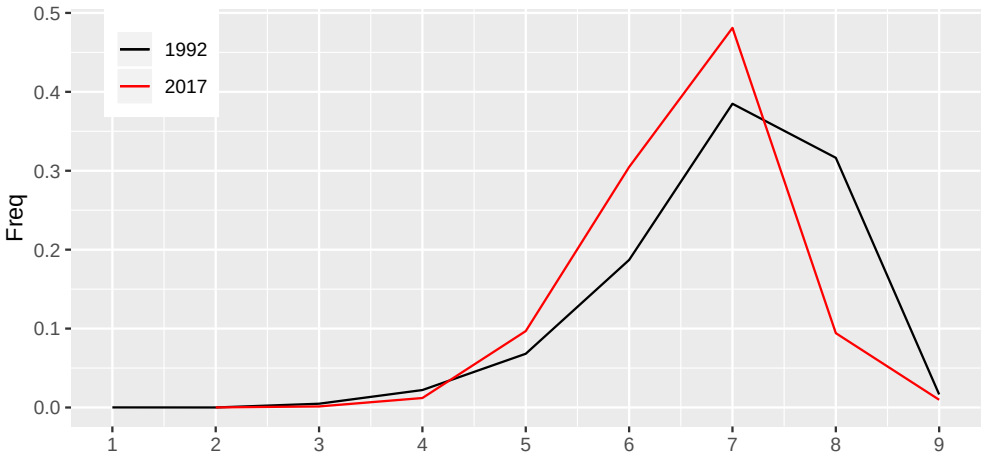

Figure 3.4: Mean condition of selected bridges (1992 vs. 2017)

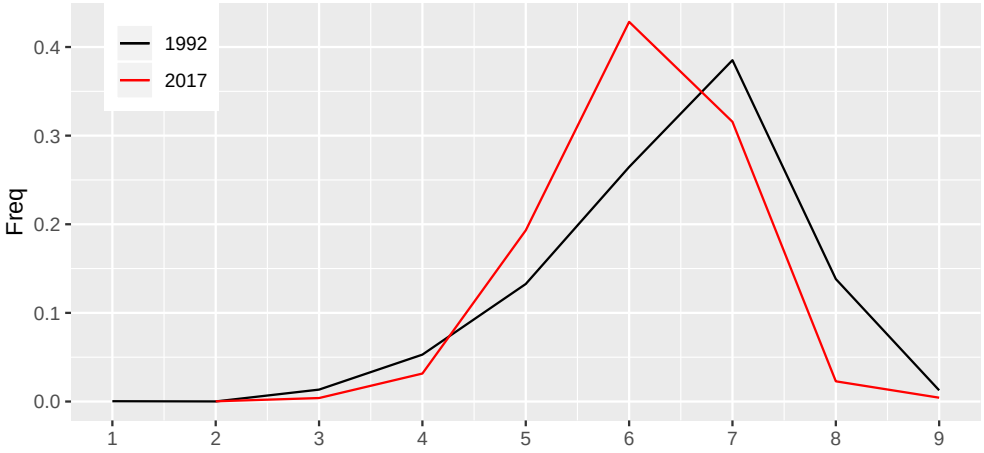

Figure 3.5: Minimum condition of selected bridges (1992 vs. 2017)

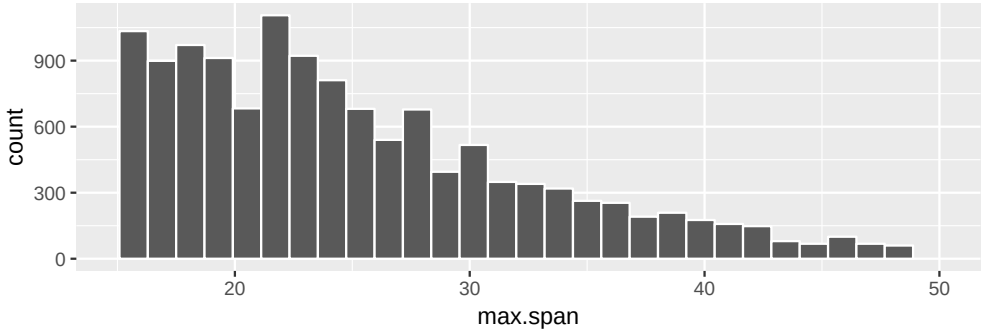

Figure 3.6: Distribution of max. span on selected bridges

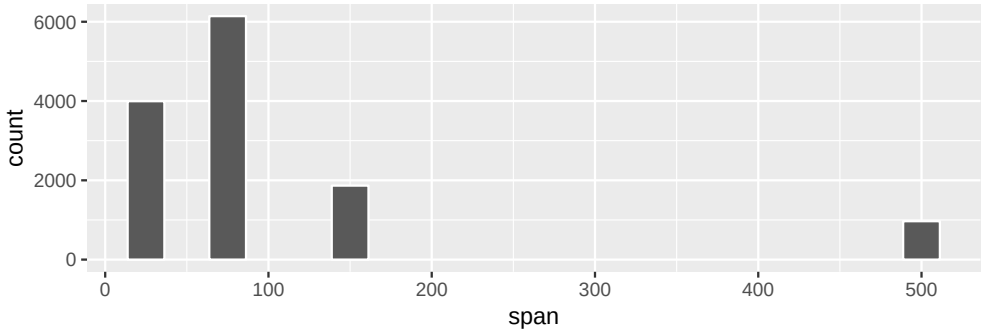

Figure 3.7: Distribution of total length on selected bridges

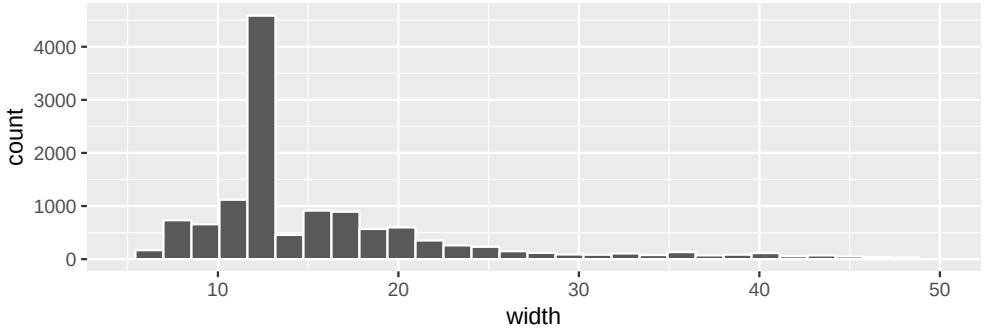

Figure 3.8: Distribution of bridge width on selected bridges

The resulting distribution of the structural condition for the time series is, on average: * Structurally deficient (rating of 4 or under): 3.86 % * At risk of being deficient (rating of 5): 15 % * Non-deficient: 81.13 %

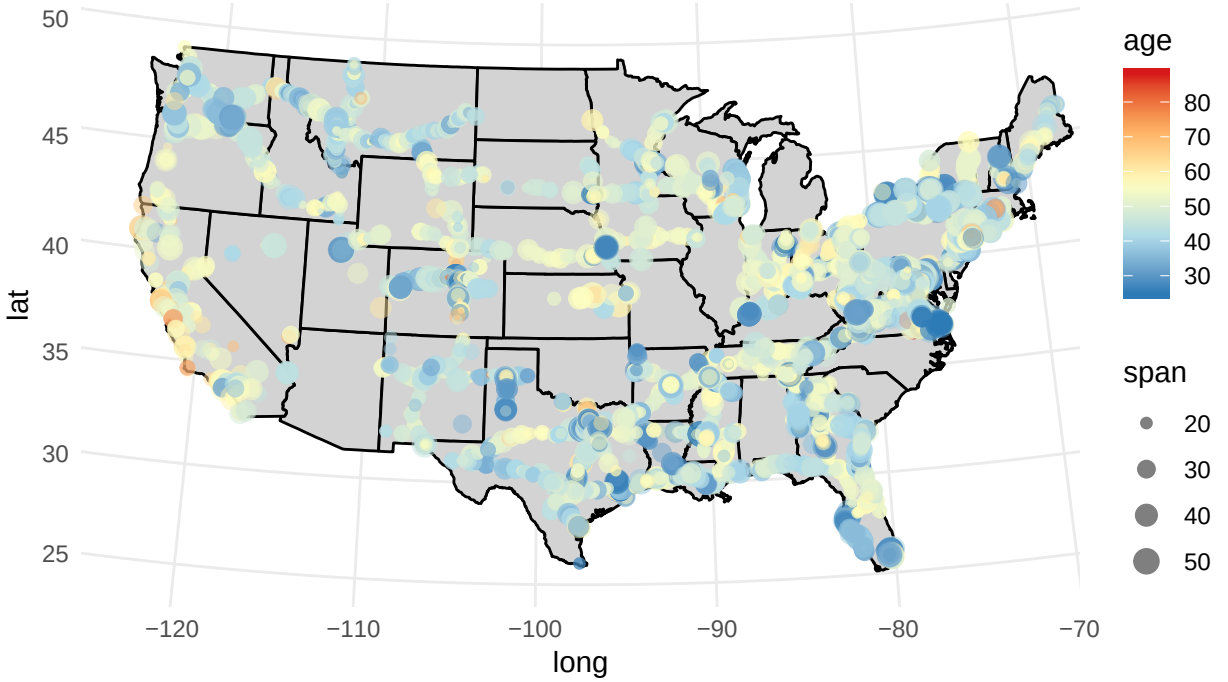

Figure 3.9: Age and span spatial distribution of selected bridges (2017)

Chapter 4

Methodology and results

4.1 Extreme Clustering

The first analysis consisted of a modification of a clustering technique - partition around medoids - used in Bracken et al. (2015).

3,000 case studies were randomly sampled from the selected data to alleviate the simulation. The structural condition time series and the lat, long position of the sample was the only variables provided. A range from 2 to 20 clusters was analyzed. Figures 4.2 and 4.3 show the optimal and third-best performance according to the average silhouette method reproduced in figure 4.1.

The similarity between the pattern showed by the 6-cluster simulation and the North American Climate zones/types made us reflect about a potential connection. That is why new variables associated with monthly extreme temperatures and annual precipitation were added as covariates.

4.1.1 Adding climate variables to bridge location

Monthly climate-division data from the National Centers for Environmental Information (Vose et al., 2014) was collected for the period 1991-2017. For each of the 344 datapoints scatterd throughout the Continental US, the anual monthly maximum and minimum temperature and annual precipitation was aggregated.

To translate this information into bridge-located data, a local polynomial regression with a low alpha (0.05) and linear degree was used for each climate variable. Fig

4.2 Principal Component Analysis

A second analysis was carried out using PCA. In this case all variables with exception of the deficiency condition of the bridge were considered in the simulation. As a first approach, the climate variables were not included (Fig. 4.6). Repeating the process with tmax, tmin, and prec lead to a greater explanation of the variance in firsts four PC (see fig. 4.7).

With the second analysis (climate - considered), the influence of each eigenvector on the attributes of our dataset was plotted and analyzed (fig. 4.8)

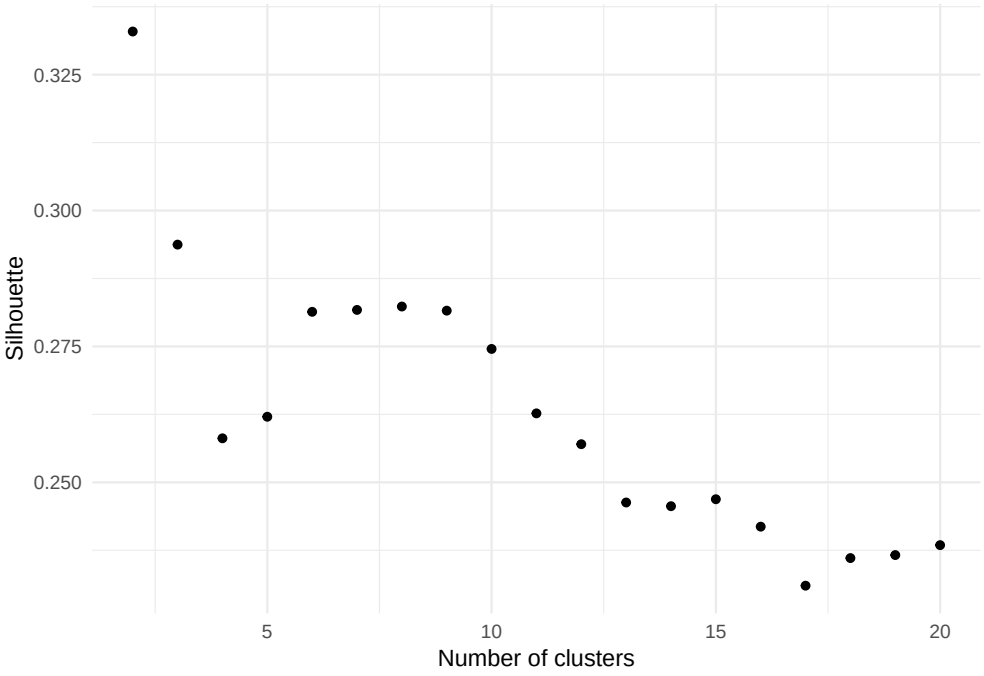

Figure 4.1: Average silhouete profile for each number of clusters

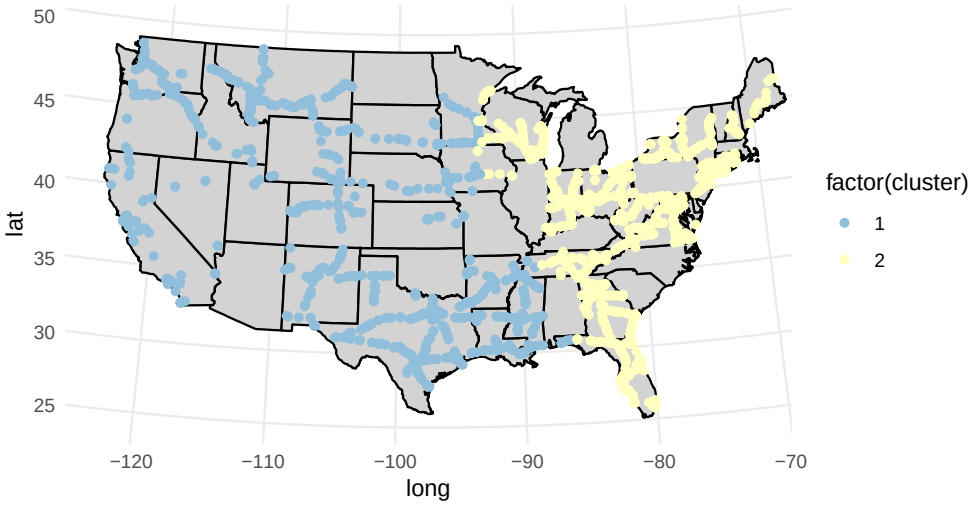

Figure 4.2: Optimal PAM 2 - Clusters

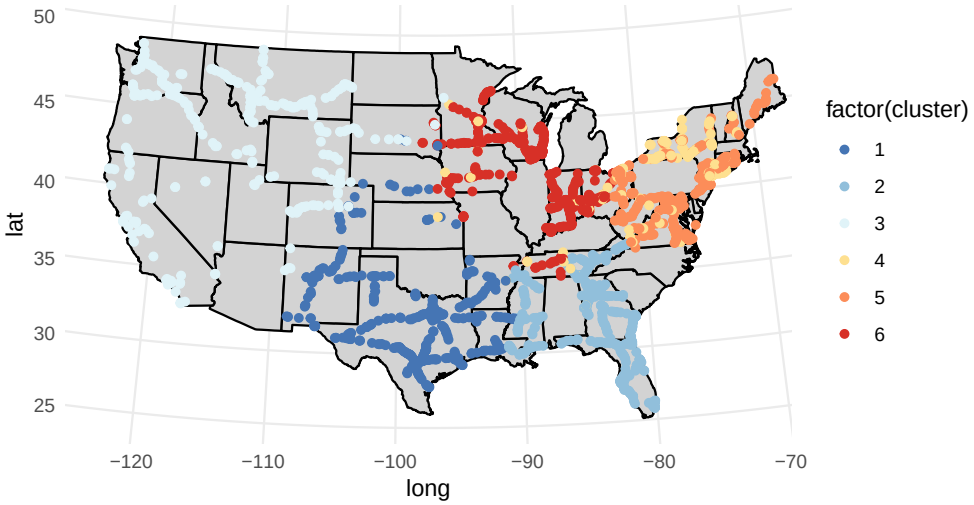

Figure 4.3: Sub-optimal PAM 5 - Clusters

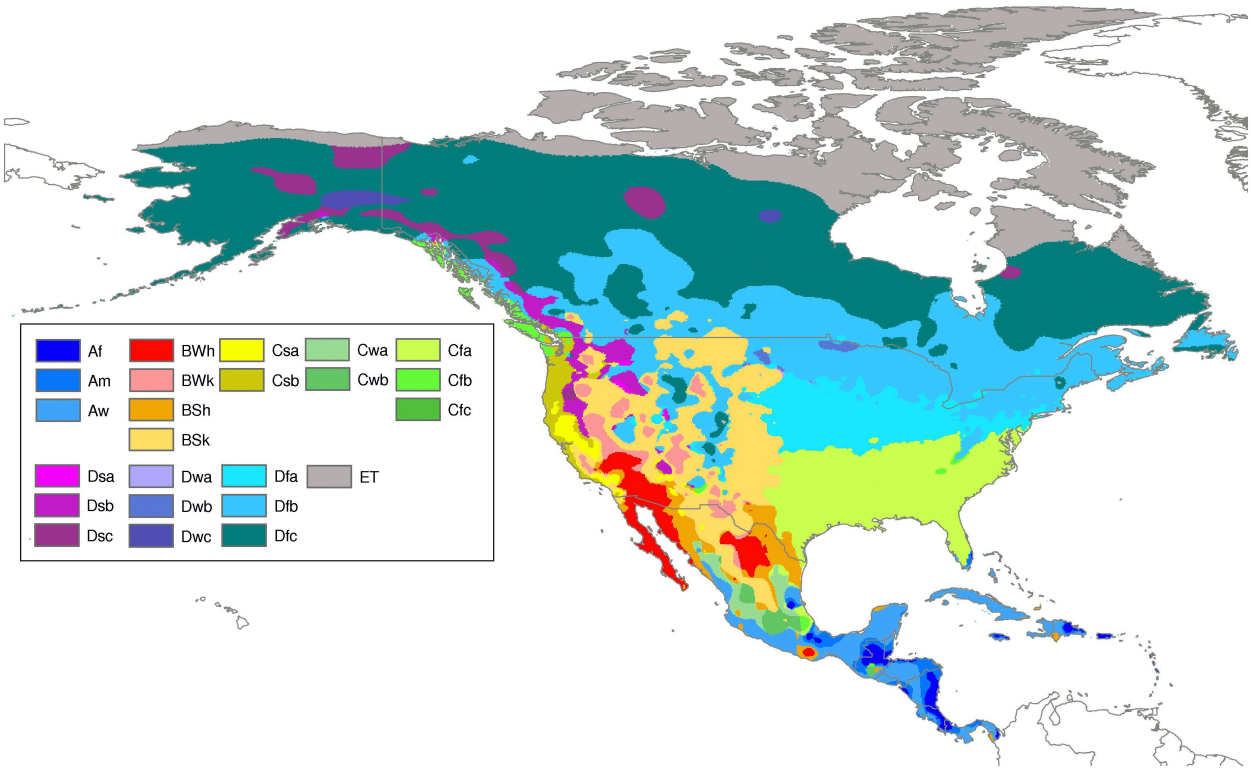

Figure 4.4: Climate types according to Koppen classification

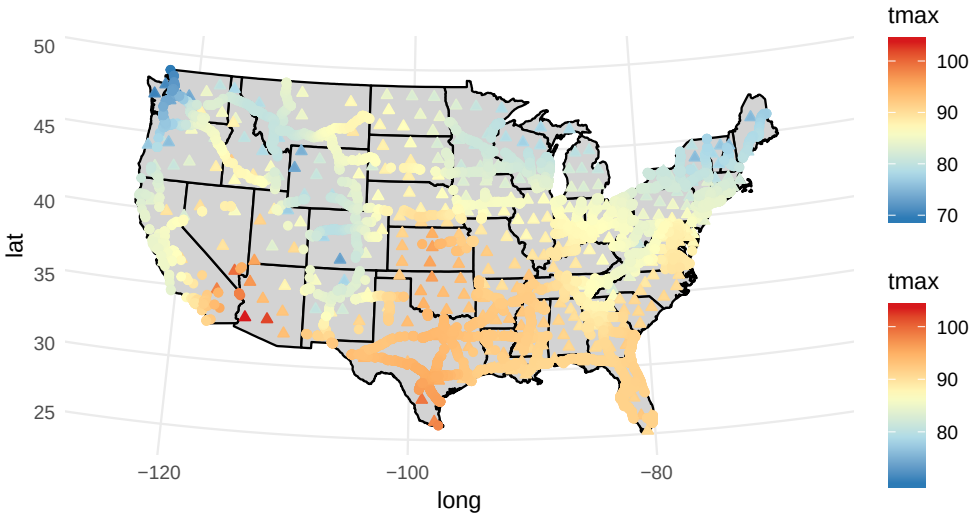

Figure 4.5: 1992 Annual maximum temperature - nClimDiv data and locfit regression

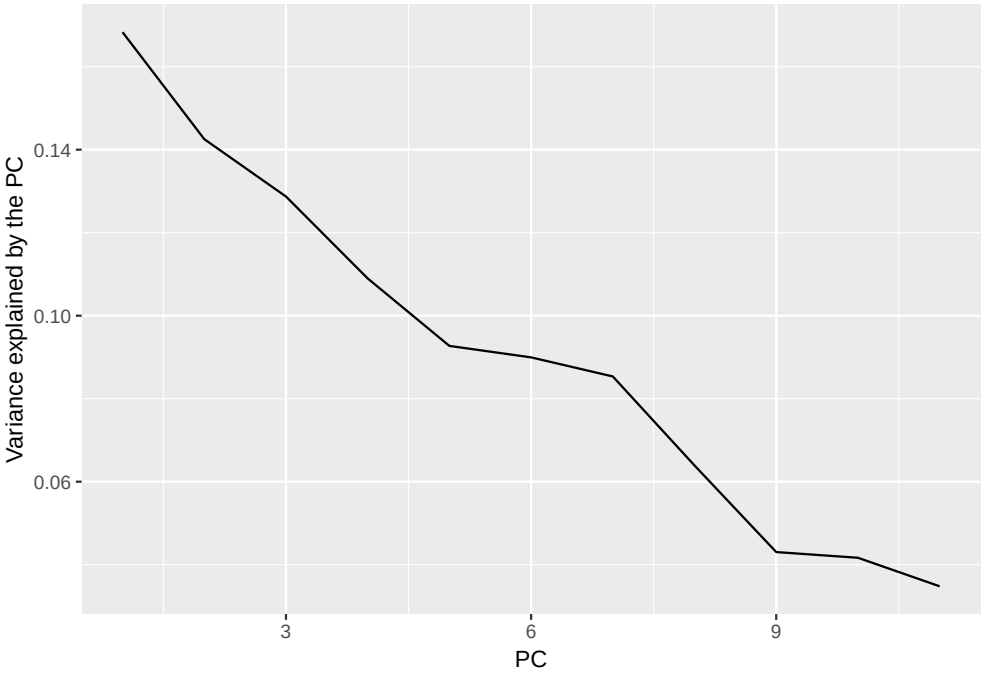

Figure 4.6: Variance explained by each PC

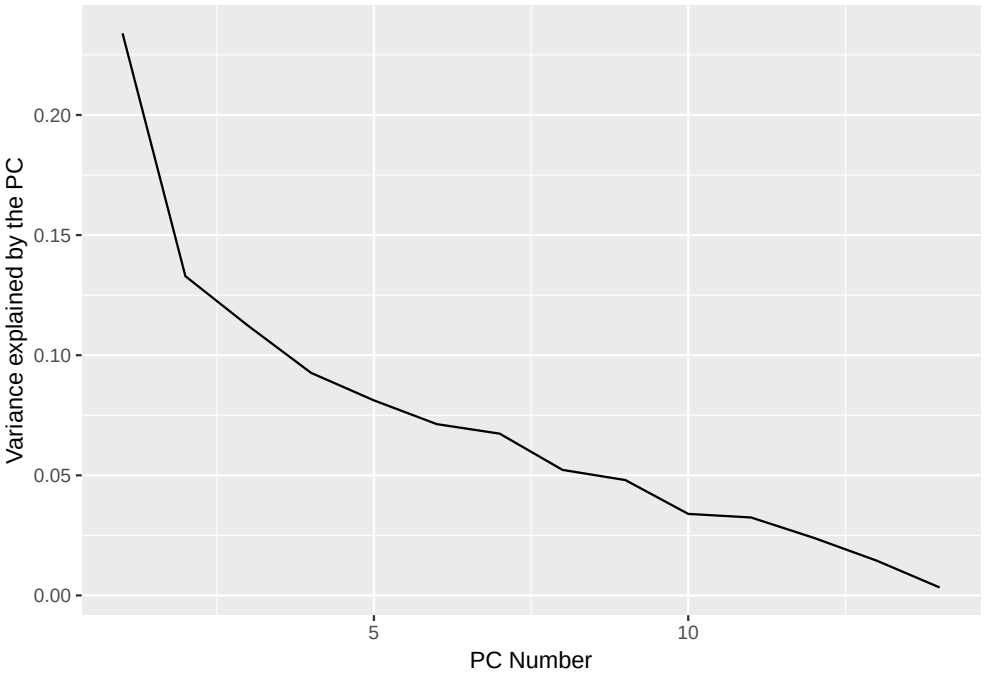

Figure 4.7: Variance explained by each PC, including climate variables

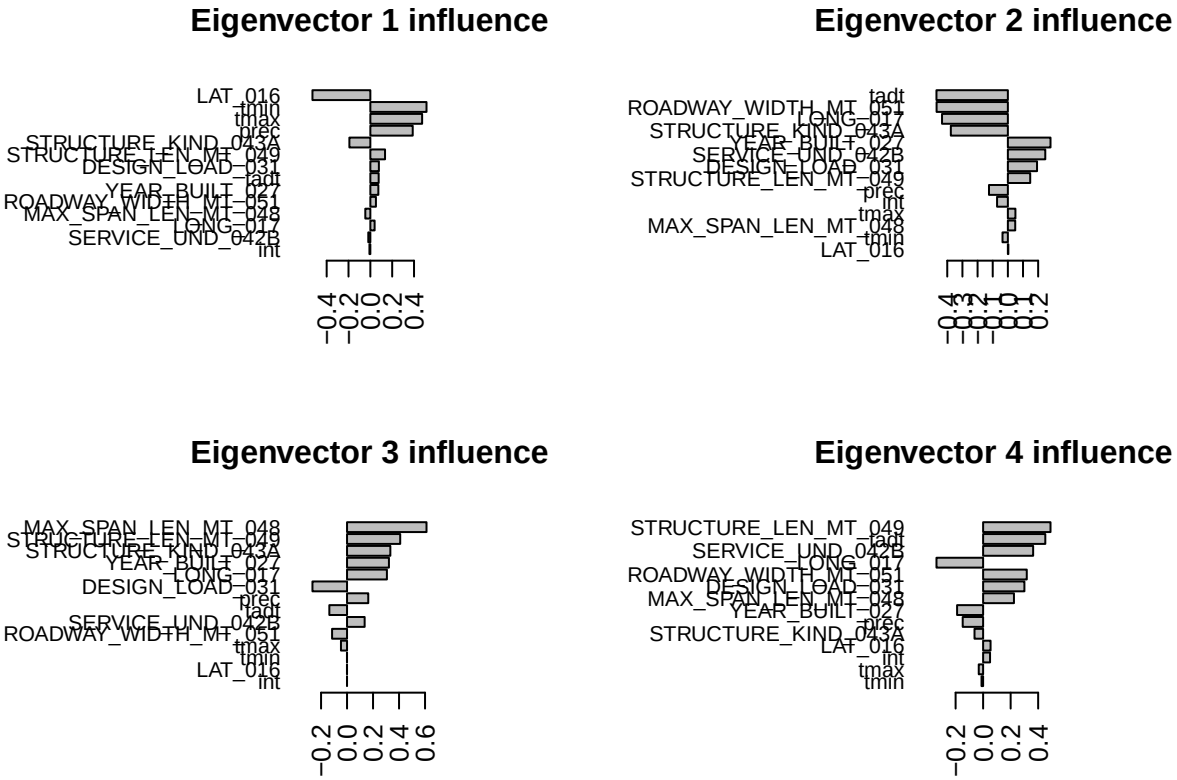

Figure 4.8: Influence of eigenvector on each attribute

4.2.1 Multinomial regression

The first 4 PC were used to fit a multinomial regression on the structural condition of the bridge, as they represent almost 60% of the variance of the model. The best model using step AIC criteria was the one including all four PC as covariates.

Two statistics were calculated to assess the accuracy of the regression. First, the ranked probability score for the model without climate variables was of 7.5 %. That is, the increase in the accuracy by predicting through the model instead of the count based probabilities was of 7.5 %. The introduction of the climate variables (tmax, tmin, prec) increased the ranked probability score skill from 7.5 % to 8.6 %.

Additionally, a confusion matrix `caret::confusionMatrix` was calculated to evaluate the "false positives" and "false negatives" the model predicted. The output shown below evidences how hard it is for the model to be accurate, as the number of "false 1 identified as such" (structurally deficient) is similar to non identified "true 1". Only a small fraction of "true 1" are identified correctly.

```
##            Reference
## Prediction    1    2    3
##          1   78  151  613
##          2  144  303 1249
##          3  633 1236 8563
```

4.3 Self Organizing Maps

As an alternative to the previous machine learning techniques, a 3 by 3 node SOM clustering has been carried out. Figure 4.9 the weight of each attribute on each of the 9 nodes. The resulting distribution aggregates statistically closer bridges in the same nodes, with a potential for better accuracy in the regression of each separately.

To better explain this approach, two properties from the ensemble, the design load knowledge and the reconstruction effect, will be plotted. Figure 4.10 shows the strong signal of the design load variable on the red node. Similarly, the comparative relevance of bridges on an specific node related to the binary variable "reconstruction" is plotted in figure 4.11.

Alternatively, a hierarchical clustering could be carried out on the map to obtain fewer number of nodes and reduce regression models. As an example, a k-means with 4 clusters on the nodes is showed in figure 4.12.

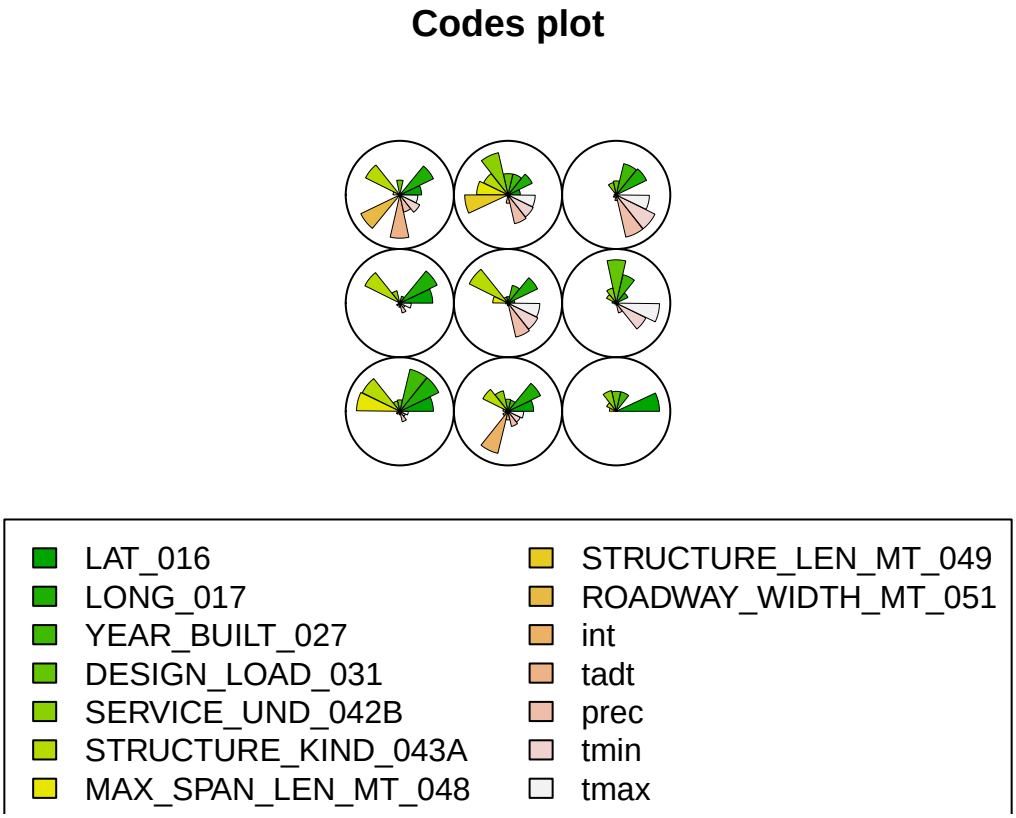

Figure 4.9: Attribute (code) signal on each node

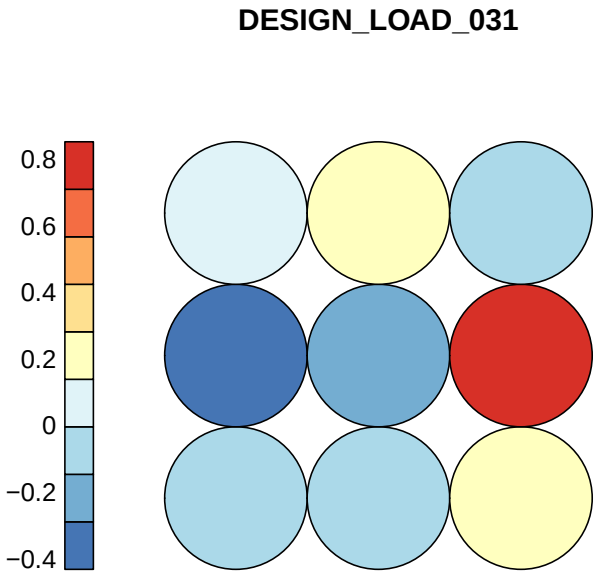

Figure 4.10: Property plot for "Design load" binary variable

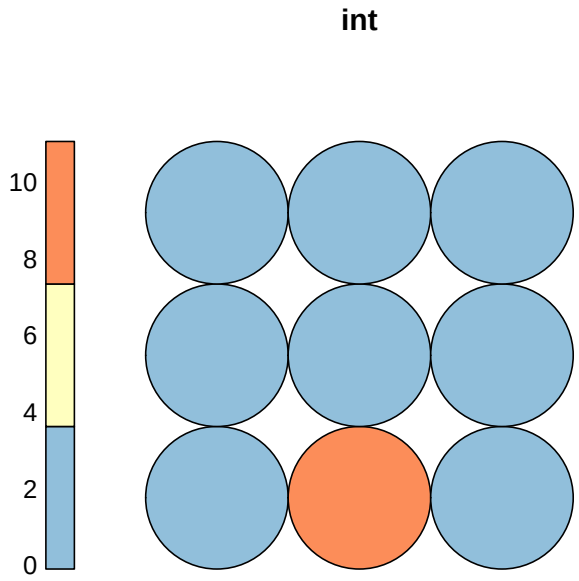

Figure 4.11: Property plot for "Reconstruction" binary variable

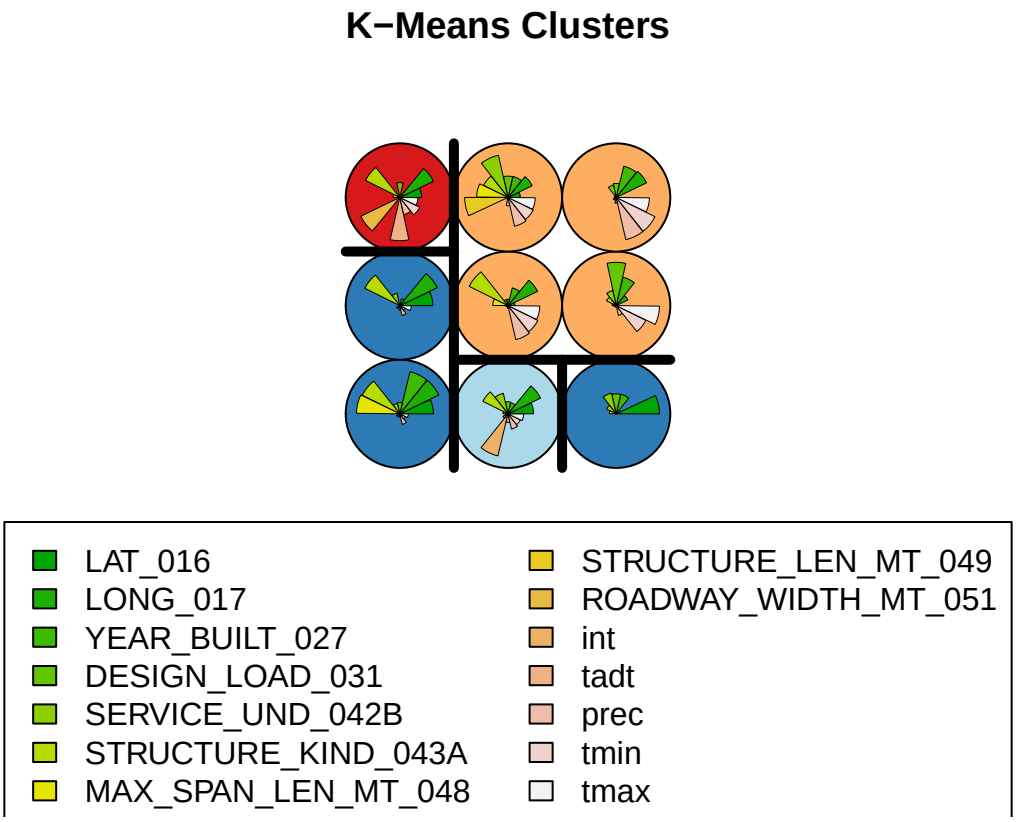

Figure 4.12: Hierarchical K-means clustering on SOM

Bibliography

ASCE (2017). The 2017 Infrastructure Report Card: A comprehensive assessment of America's infrastructure. American Society of Civil Engineers Washington, DC.

Bektaş, B. A. (2017). Use of Recursive Partitioning to Predict National Bridge Inventory Condition Ratings from National Bridge Elements Condition Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2612:29–38.

Bektas Basak Aldemir, Carriquiry Alicia, and Smadi Omar (2013). Using Classification Trees for Predicting National Bridge Inventory Condition Ratings. *Journal of Infrastructure Systems*, 19(4):425–433.

Bracken, C., Rajagopalan, B., Alexander, M., and Gangopadhyay, S. (2015). Spatial variability of seasonal extreme precipitation in the western United States. *Journal of Geophysical Research: Atmospheres*, 120(10):4522–4533.

Contreras-Nieto, C. (2014). *Development of Linear Models to Predict Superstructure Ratings of Steel and Prestressed Concrete Bridges*. PhD Thesis, Oklahoma State University.

Contreras-Nieto, C., Shan, Y., and Lewis, P. (2018). Characterization of Steel Bridge Superstructure Deterioration through Data Mining Techniques. *Journal of Performance of Constructed Facilities*, 32(5):04018062.

Saeed, T. U., Moomen, M., Ahmed, A., Murillo-Hoyos, J., Volovski, M., and Labi, S. (2017). Performance Evaluation and Life Prediction of Highway Concrete Bridge Superstructure across Design Types. *Journal of Performance of Constructed Facilities*, 31(5):04017052.

Veshosky, D., Beidleman, C. R., Buetow, G. W., and Demir, M. (1994). Comparative analysis of bridge superstructure deterioration. *Journal of Structural Engineering*, 120(7):2123–2136.

Vose, R. S., Applequist, S., Squires, M., Durre, I., Menne, M. J., Williams, C. N., Fenimore, C., Gleason, K., and Arndt, D. (2014). Improved Historical Temperature and Precipitation Time Series for U.S. Climate Divisions. *J. Appl. Meteor. Climatol.*, 53(5):1232–1251.

Weseman, W. A. (1995). Recording and coding guide for the structure inventory and appraisal of the nation's bridges. *United States Department of Transportation (Ed.), Federal Highway Administration, USA*, 119.