STAT 503 Homework 4: STAT 101 Grades

Andee Kaplan, Will Landau, Fangge Liu, Lindsay Rutter

March 27, 2015

Introduction

STAT 101 course instructors at Iowa State University usually claim their students are diverse. The undergrads who sign up have a wide variety of majors, backgrounds, perspectives, abilities, and levels of motivation. Visual and unsupervised analyses of homework grades may classify students in useful, insightful ways and even inform pedagogy.

We have three homework grade spreadsheets, each of which comes directly from Blackboard (Iowa State University 2015), Iowa State's system for managing course materials and grades. Each dataset corresponds to a single semester of STAT 101: either fall 2013, fall 2014, or spring 2014. Each semester has six or seven sections of roughly one hundred students each. Every spreadsheet has roughly twenty variables, each of which corresponds to a homework grade (either percentage of points earned or NA for a missing assignment) or the average homework score with missing assignments removed.

Missing values

A large fraction of homework scores appear as "NA", or missing. Since the spreadsheets come directly from Blackboard, we assume that almost all NA's correspond to homeworks that students failed to turn in, and only a small number, if any, are from bookkeeping errors. Before clustering, we need to impute these values, and for an imputation strategy, we look at patterns of missingness.

Visual patterns in missing values

Figure 1 plots the number of missing values per student for all students. Most of the students missed few to no assignments, and some students missed several. This pattern is consistent with each semester and section. Figure 2 shows the number of missing values for each assignment within each semester and section. For the fall semesters, notable spikes occur at chapter 9 ("Understanding Randomness") and topic 9 ("Sample Surveys"). Otherwise, for most sections, the number of missings increased steadily for each section over time.

Figures 3, 4, and 5 show each student's pattern of missingness. General patterns are consistent. Most students had few missing assignments scattered sporadically over the semester, some students dropped the class early, and a smaller students missed strings of around five or ten consecutive assignments in the beginning, middle, or end of the semester.

Grouping students by missingness

From inspection, we can partition the students in each semester into four groups.

Group 1 contains students who did not submit all of the last nine (about half) of all homework assignments. Group 2 contains students (who were not in Group 1) and missed at least nine (about half) of all homework assignments. Group 3 contains students (who were not in Group 1 or 2) and missed at least one homework assignment. Group 4 contains students (who were not in Group 1, 2, or 3) who did not miss any homework assignments.

As a result, we generated 12 main groups (the four groups across three semesters), as shown in Table 1.

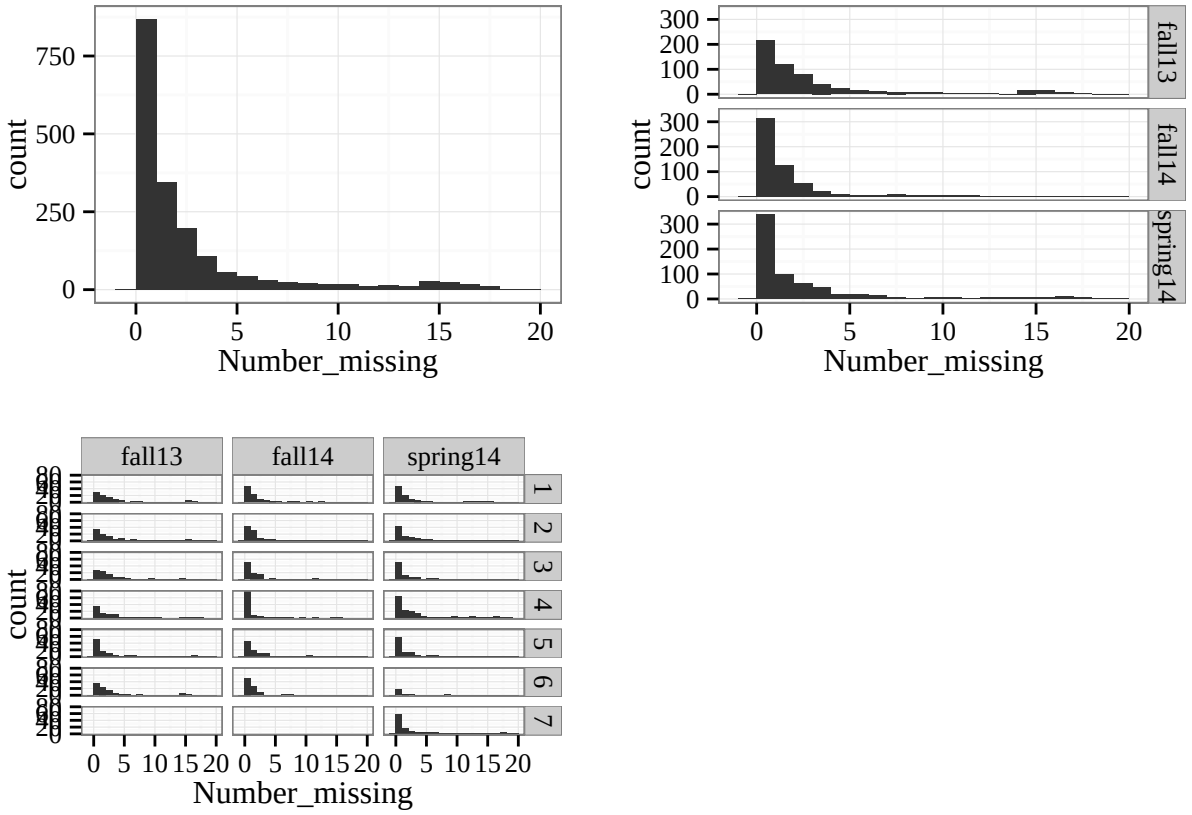

Figure 1: number of missing values per student for all students. The top right panel facets by semester, and the bottom left panel facets by semester and section number. Most of the students missed few to zero assignments, and some students missed several. This pattern is consistent with each semester and section.

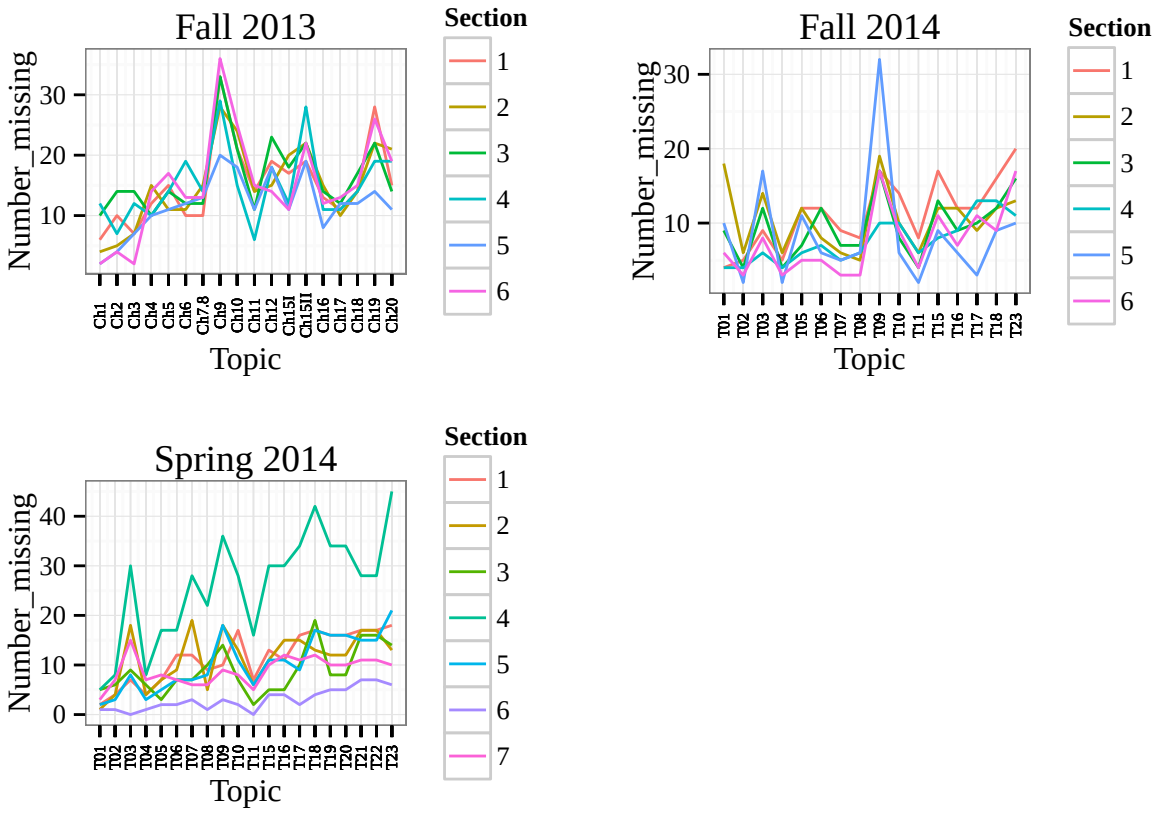

Figure 2: number of missing values per assignment for each semester and section. For the fall semesters, notable spikes occur at chapter 9 ("Understanding Randomness") and topic 9 ("Sample Surveys"). Otherwise, for most sections, the number of missings increased steadily for each section over time.

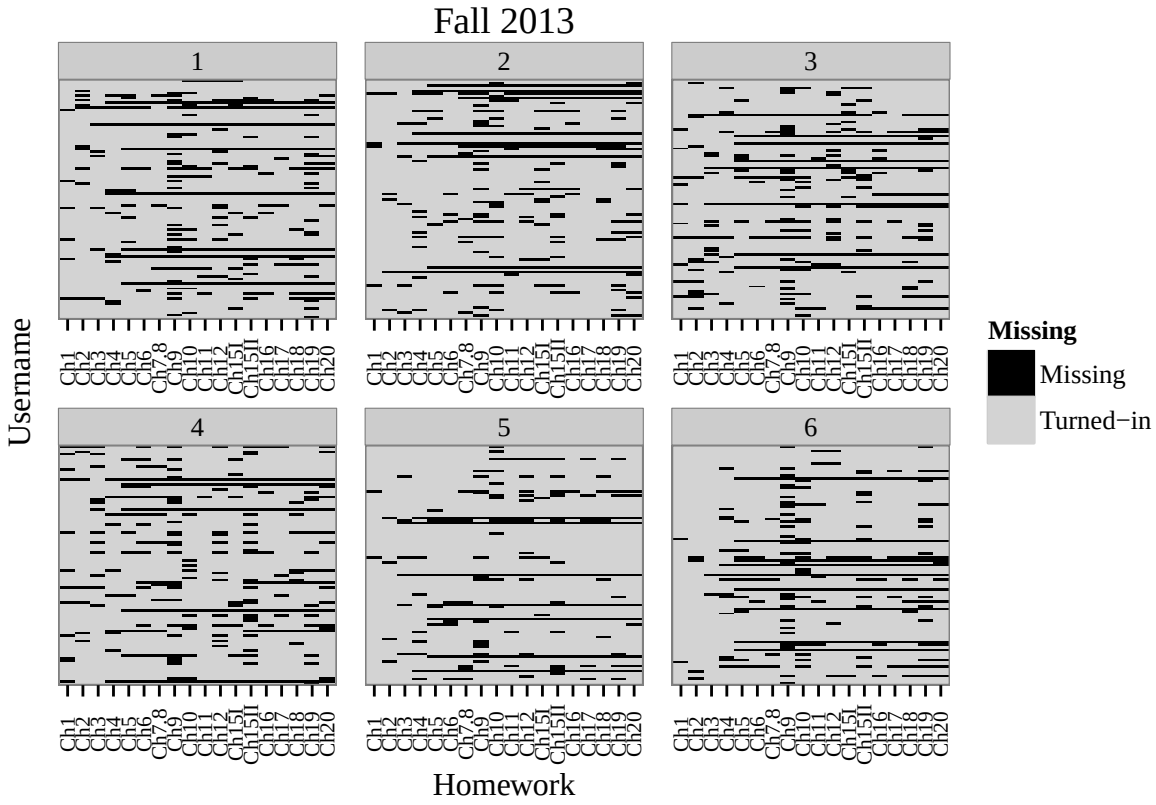

Figure 3: missing assignment records for fall 2013 students faceted by section number. Each row represents a student, each column is an assignment, and the tiles are colored according to the status of the corresponding assignment (missing or turned in). General patterns are consistent across section number. Most students missed few assignments, and the students with the most missings usually missed the last fifteen assignments. These students most likely dropped the class early.

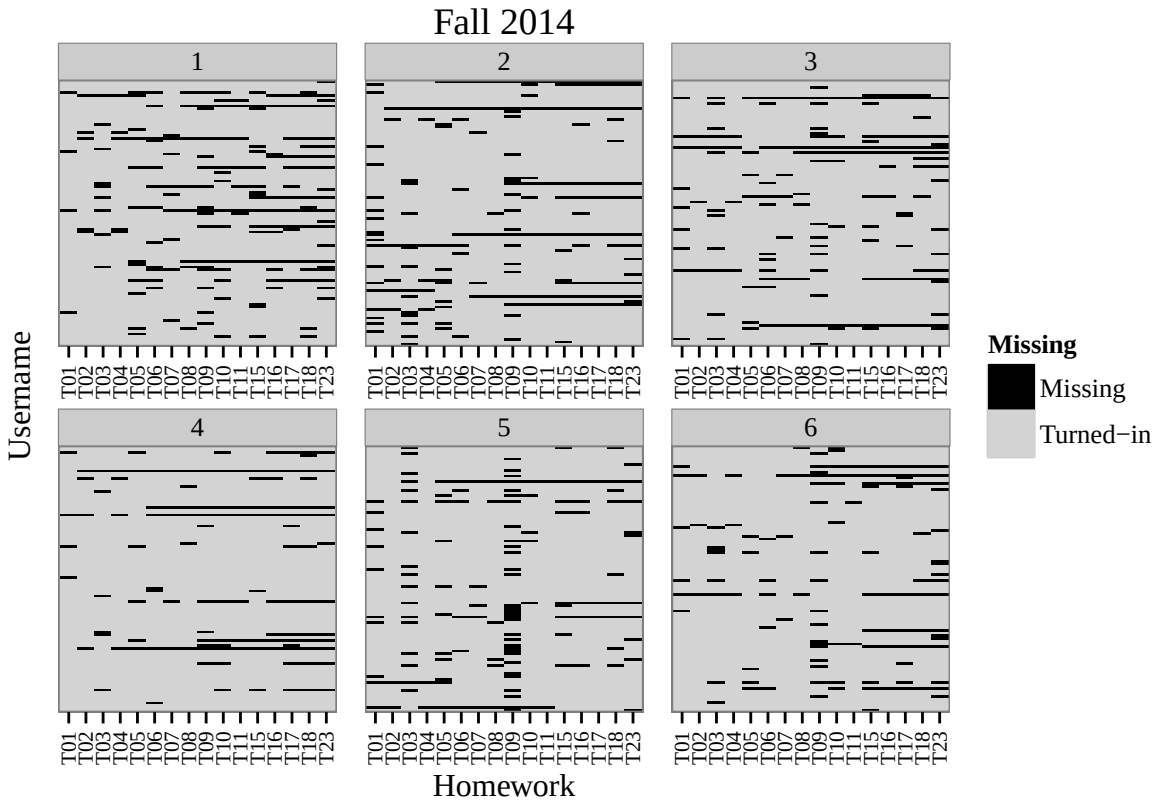

Figure 4: Same as Figure 3, but for fall 2014. We see some early drops, but also some students who failed to turn in either the first few or the middle few assignments.

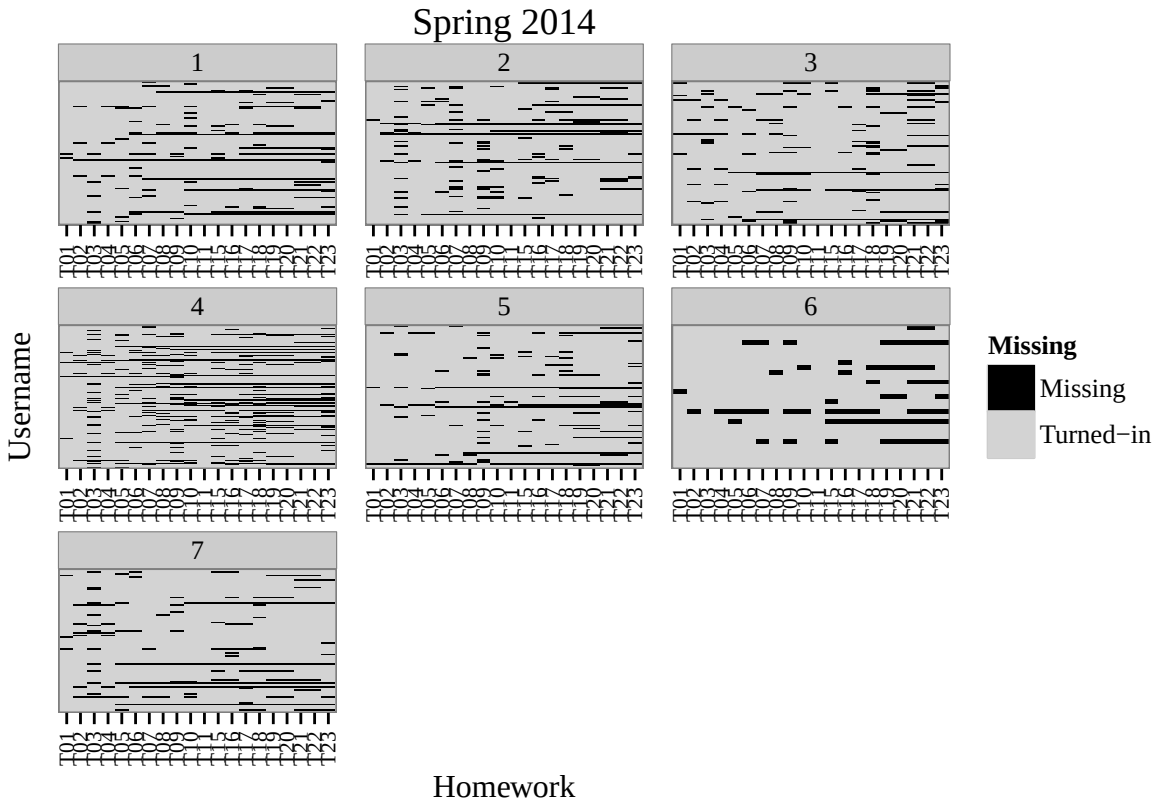

Figure 5: Same as Figure 3, but for spring 2014. Patterns are similar to those of Figure 3 (fall 2013).

Table 1: The number of students who were categorized into one of four mutually-exclusive groups, for the three semesters. We considered Groups 1 and 2 to be problematic, as they likely represented students who dropped the course or habitually missed assignments. However, even after removing students from Groups 1 and 2, we are still left with a very large dataset to cluster

	Fall 13	Spring 14	Fall 14
Group 1 - Drop outs	45	48	17
Group 2 - Common missings	24	23	20
Group 3 - Sporadic missings	298	268	236
Group 4 - No missings	216	338	314

Trim and impute

As we explain above, groups 1 and 2 are relatively small, and the students in these groups missed at least half of the homework assignments. With good reason, we give each of these groups its own cluster and concentrate the rest of our analysis on groups 3 and 4 only. These remaining students have some missing values left, and we impute them with the nearest neighbors imputation functionality in the DMwR (Torgo 2010) package. Dr. Cook wrote most of this code.

A look at the cleaned data

The parallel coordinate plot in the problem statement shows most students tended to do fairly well on homework, and Chapters 1, 9, 12, and 19 may be nuisance variables because of the lack of variability. Figure 6 shows histograms of all the homeworks scores (and the average score) for fall 2013. The results for other semesters are similar. It will be difficult to split on any one variable individually except for maybe chapter 12. Chapters 1 and 2, along with the average score, look like nuisance variables due to low variability.

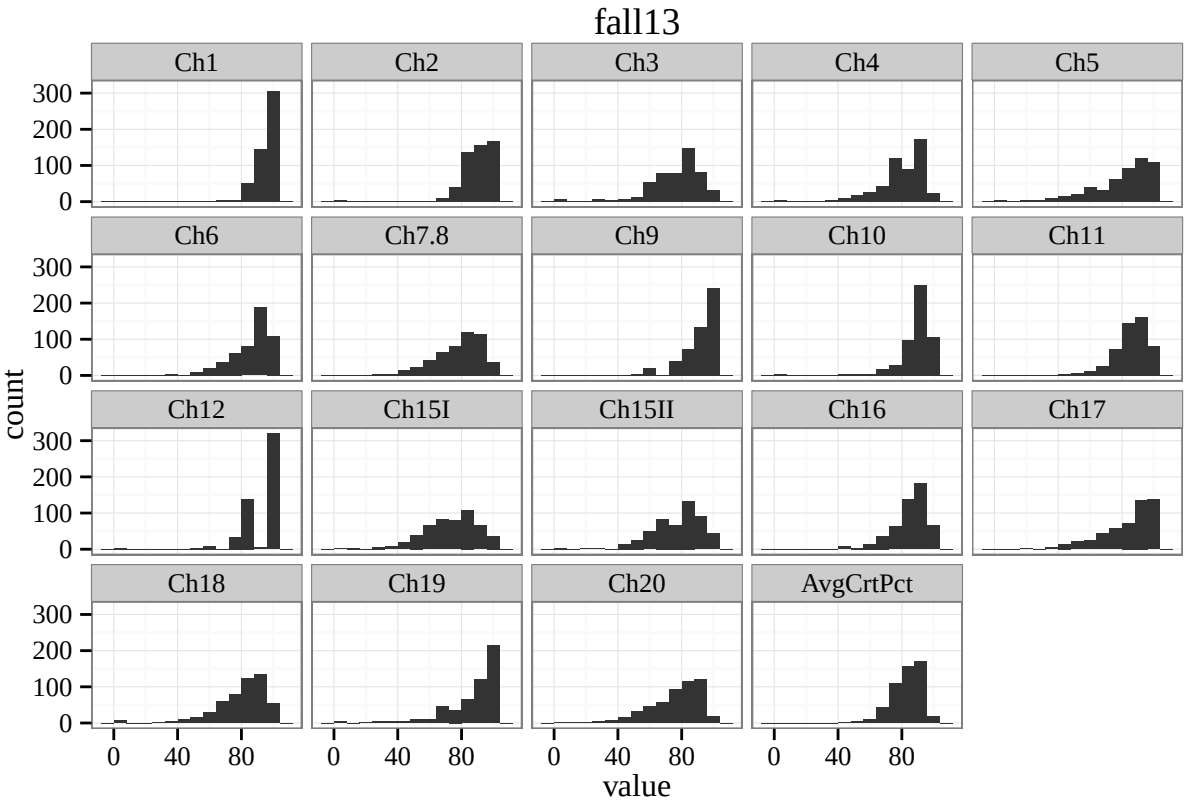

Figure 6: histograms of all the homeworks scores (and the average score) for fall 2013. The results for other semesters are similar. It will be difficult to split on any one variable individually except for maybe chapter 12. Chapters 1 and 2, along with the average score, look like nuisance variables due to low variability.

Figure 7 shows a scatterplot matrix of the first ten variables from fall 2013. The results of the full 20 variables are similar. The It will be difficult to split the data on any pair of variables, as no pair provides much of a spatial separation in the data. In addition, we have a lot of nuisance variables: Figure 8 shows that the homework scores are often highly correlated.

The dimensionality of the data makes visualization difficult, so Figure 9 shows the first two principal components for each semester, and Figure 10 plots the dimensionality reduction from multidimensional scaling to two dimensions. In both cases, the data do not spatially separate into clusters. Geometrically, each semester's grades coagulate into a blob. It will be difficult, if not impossible, for typical hard clustering techniques, or any other techniques that assume an obvious visual partitioning, to make any headway. The next section briefly attempts these techniques anyway, self-organizing maps and model-based clustering are much more promising directions.

Figure 10 plots the dimensionality reduction from multidimensional scaling to two dimensions.

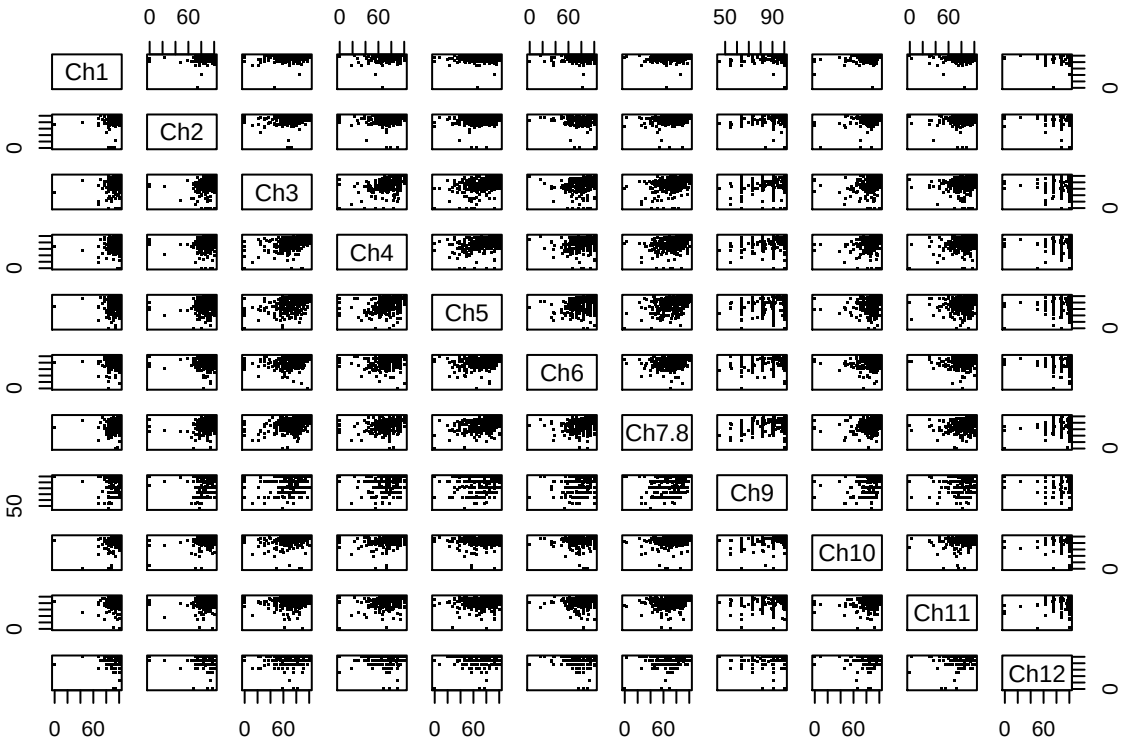

Figure 7: scatterplot matrix of the first ten variables from fall 2013. The results of the full 20 variables are similar. It will be difficult to split the data on any pair of variables, as no pair provides much of a spatial separation in the data.

Attempts at typical hard clustering

Here, we try some hard clustering techniques that assume the data separate into clusters spatially, just in case there is a hidden spatial separation that we didn't detect in the previous section. At first, we see if wb.ratio gives us a quick answer for the number of clusters in the data. Figure 11 shows wb.ratio as a function of k, the number of clusters. We show results for kmeans with Euclidian distance (with the kmeans function in core R), kmeans with correlation distance (in the amap package (Lucas 2014)), and hierarchical clustering with several linkage methods (with the hclust function in core R). It is alarming that wb.ratio does not decrease monotonically with k. Increasing the number of clusters should improve clustering outcomes. Kmeans with correlation distance is the closest to having monotone decreasing wb.ratio, but it also has the overall highest wb.ratio. This is one indication that the data may not separate. Typical hard clustering just may not work.

We also look at dendrograms from hierarchical clustering to get a sense an optimal k, if we can determine k at all. Figure 12 shows dendrograms from hierarchical clustering using six linkage methods. Only spring 2014 dendrograms are shown, as results for the other two semesters are similar. The ward linkage dendrograms indicate that $k = 3$ may be reasonable, and the other methods fail to find any meaningful separation.

As it turns out, ward linkage at $k = 3$ gives only a vague separation. Figure 13 shows that ward divides students into high, medium, and low performance, but that there's still a ton of variety within clusters.

Overall, there is little natural spatial separation in the data, and classical hard clustering methods tell a brief, uninteresting story. Self organizing maps and model based clustering may have more to say.

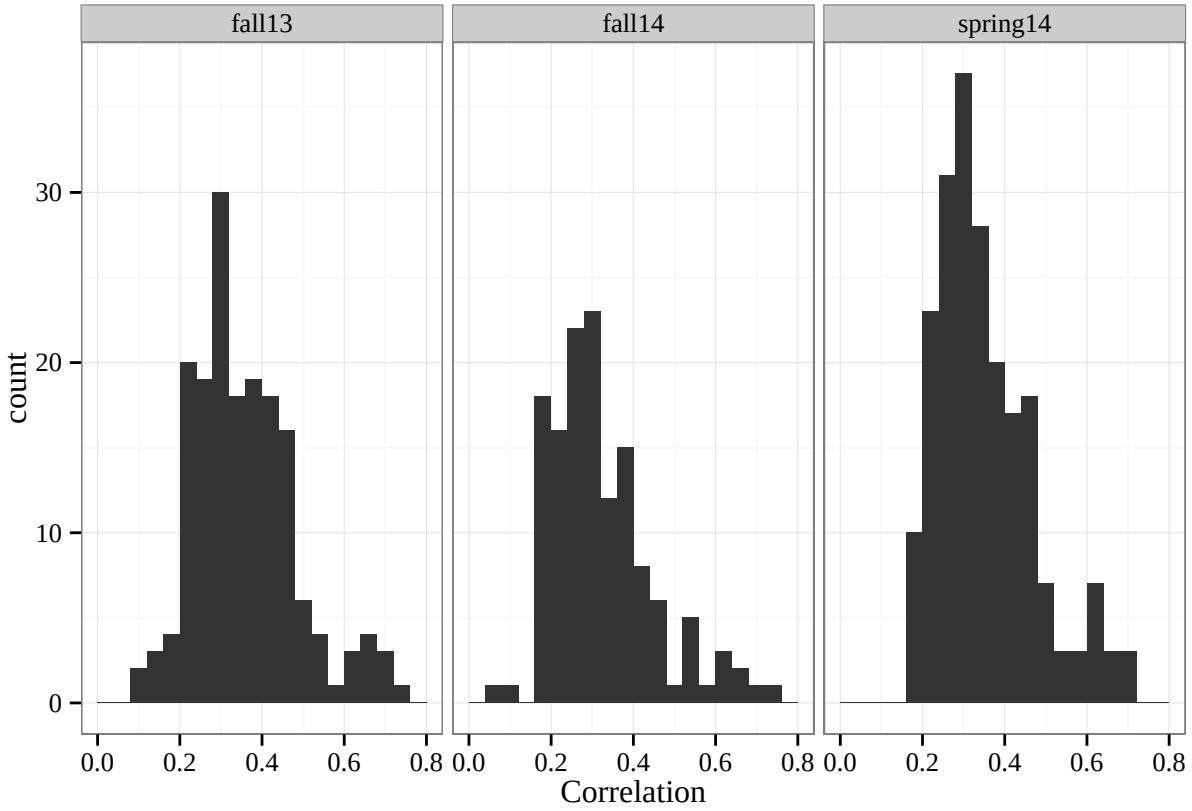

Figure 8: Histograms of correlations among homework scores, facetted by semester. The homework scores are often highly correlated, leading to nuisance variables and causing problems for typical hard clustering methods.

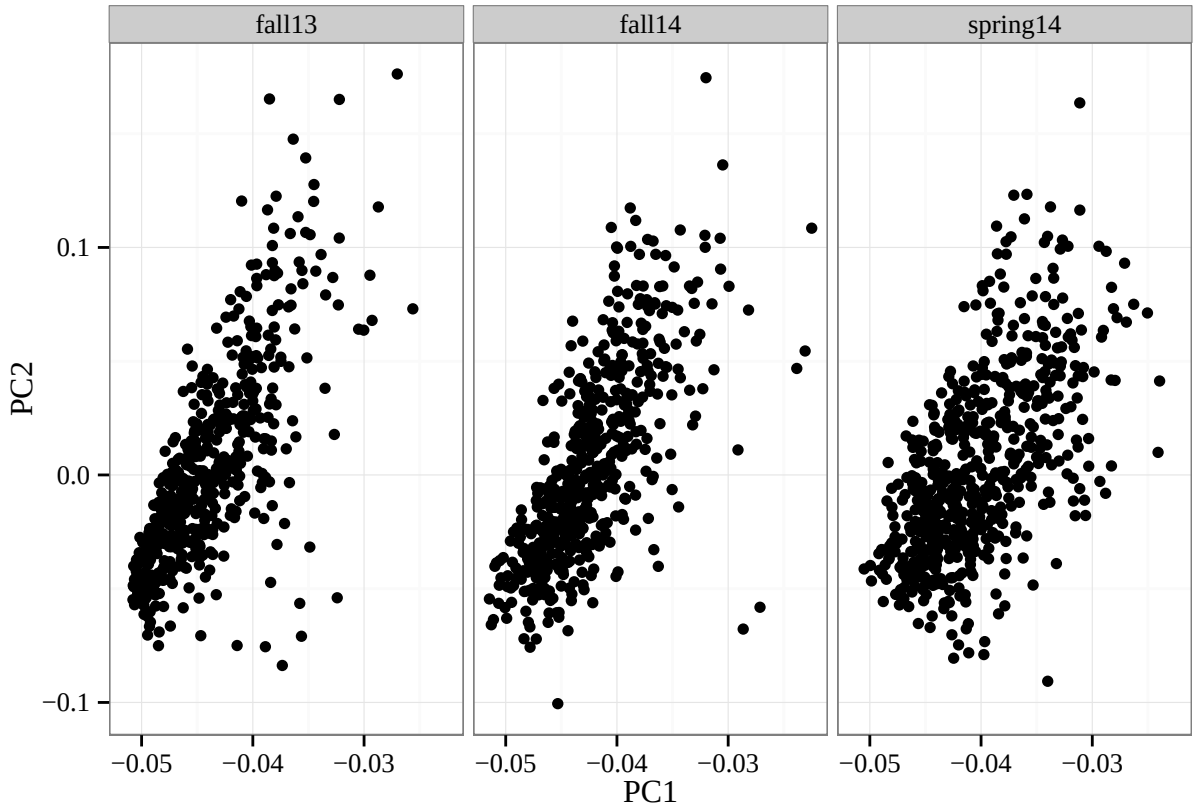

Figure 9: first two principal components, plotted against each other, for the all semesters. The data do not spatially separate.

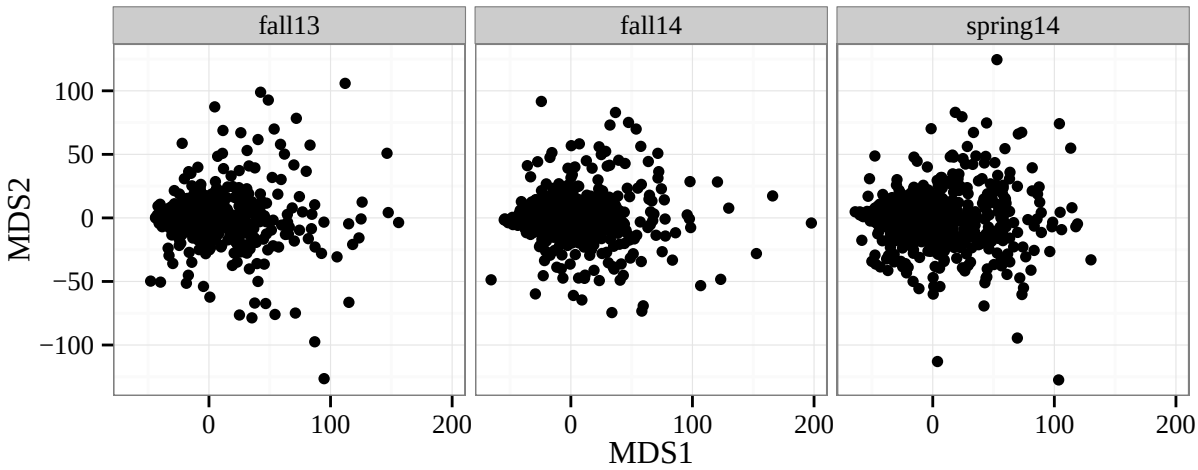

Figure 10: first two variables from multidimensional scaling, plotted against each other, for all semesters. The data do not spatially separate.

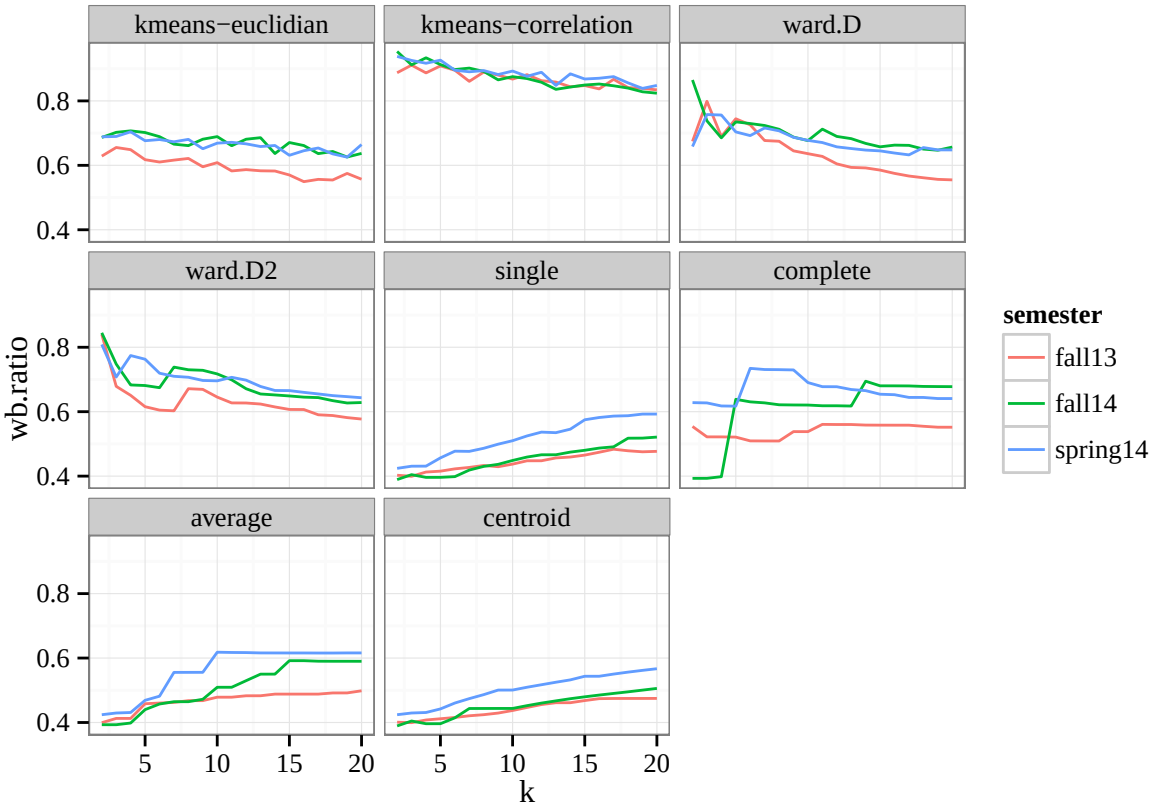

Figure 11: wb.ratio as a function of k, the number of clusters. We show results for kmeans with Euclidian distance (with the kmeans function in core R), kmeans with correlation distance (in the amap package (Lucas 2014)), and hierarchical clustering with several linkage methods (with the hclust function in core R). It is alarming that wb.ratio does not decrease monotonically with k. Increasing the number of clusters should improve clustering outcomes. Kmeans with correlation distance is the closest to having monotone decreasing wb.ratio, but it also has the overall highest wb.ratio.

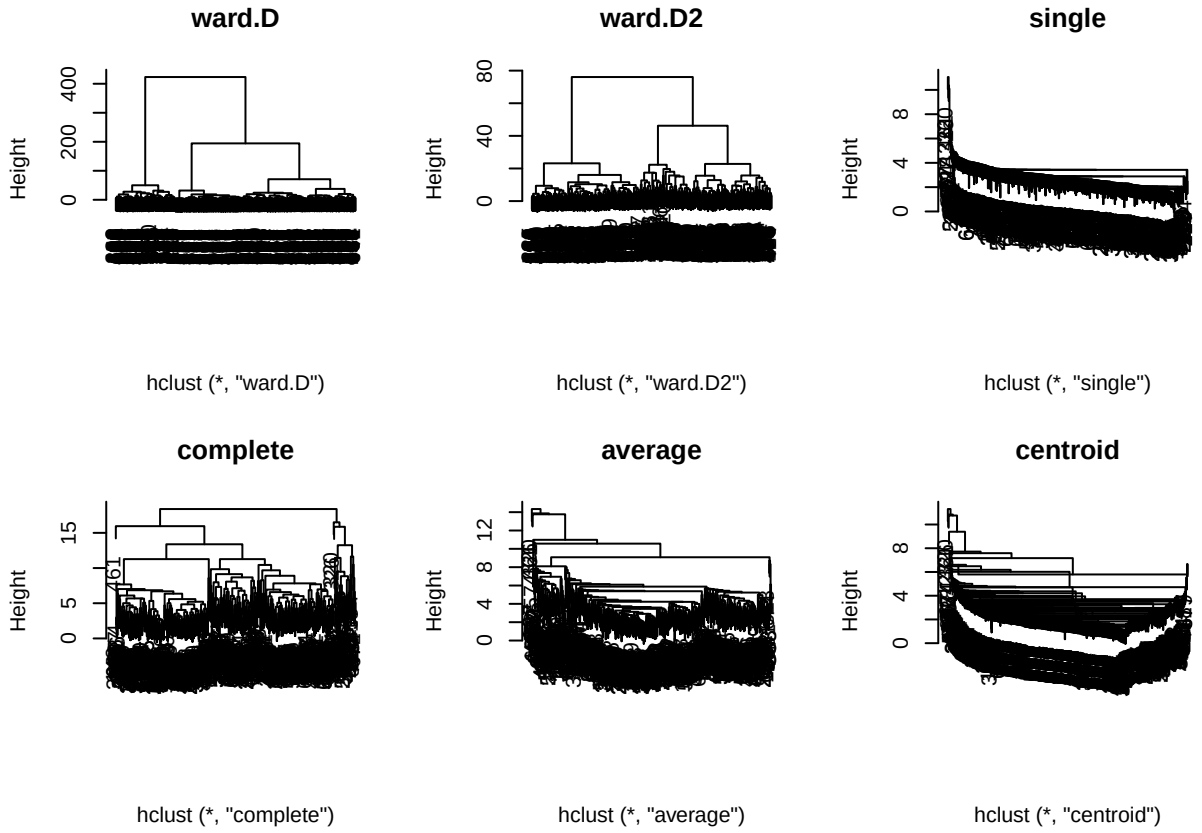

Figure 12: dendrograms from hierarchical clustering to get a sense an optimal k, if we can determine k at all. Figure ef{fig:dendros} shows dendrograms from hierarchical clustering using six linkage methods. Only spring 2014 dendrograms are shown, as results for the other two semesters are similar. The ward linkage dendrograms indicate that $k = 3$ may be reasonable, and the other methods fail to find any meaningful separation.

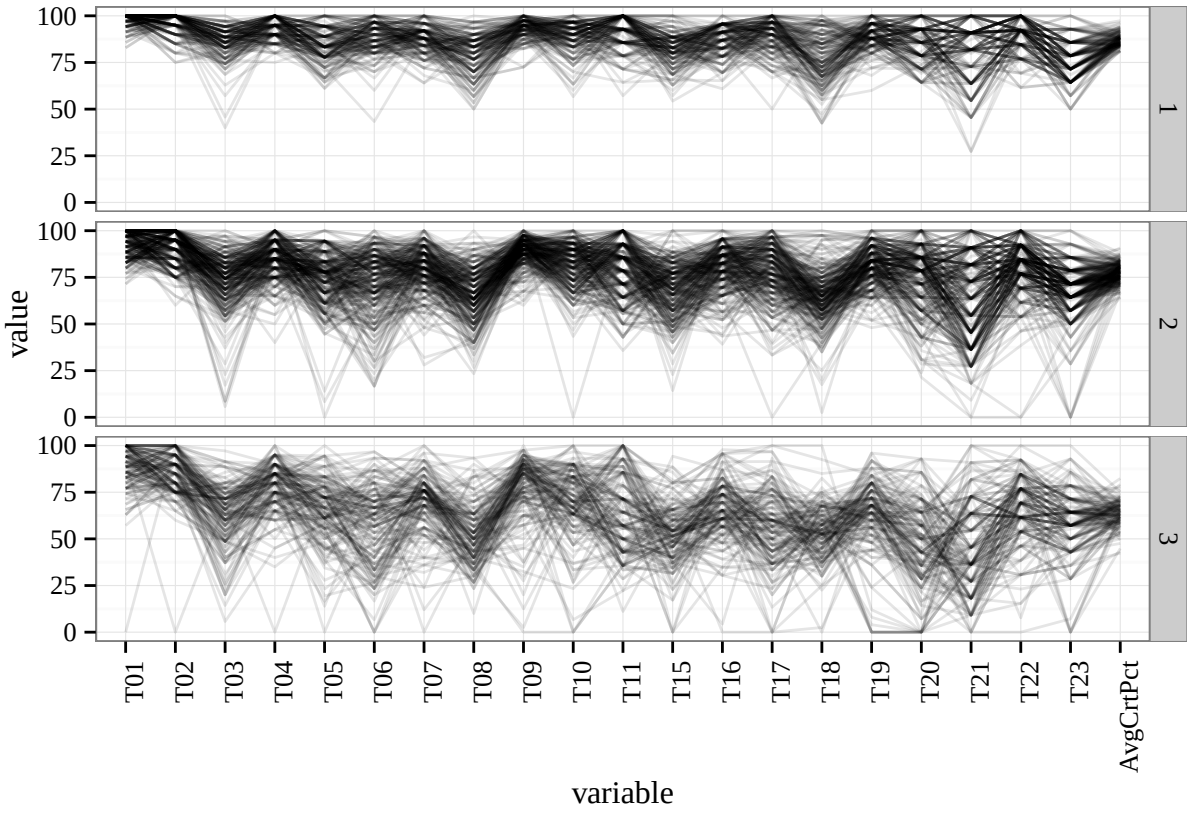

Figure 13: parallel coordinate plot of the spring 2014 scores, facetted by cluster from hierarchical clustering with ward linkage. Ward divides students into high, medium, and low performance, but that there's still a ton of variety within clusters. Overall, there is little natural spatial separation in the data, and classical hard clustering methods tell a brief, uninteresting story. Self organizing maps and model based clustering may have more to say.

Self Organizing Maps

As another attempt to cluster the students by homework grades, we fit multiple self-organizing maps (SOM) using the `kohonen` package (Wehrens and Buydens 2007) and examined the `wb.ratio` to determine how many nodes should be in the SOM grid. We examined grids of size 2×1 up to 15×15 and visualized the `wb.ratio` by grid, as well as by total number of nodes in figures 14 and 15. Each model is fit using 2000 iterations.

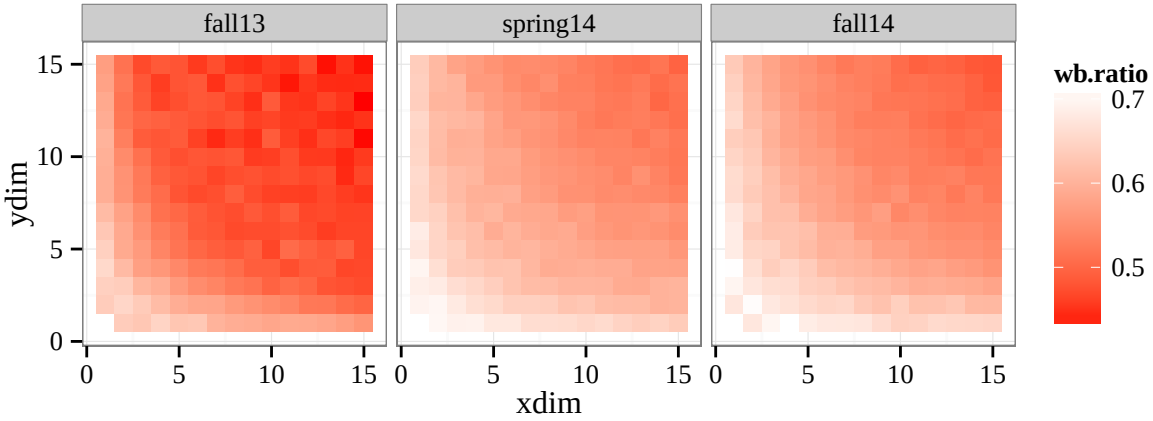

Figure 14: A plot of `wb.ratio` by grid values in the self organizing maps that were fit to homework grades. We focus on fall 2013 in order to create a map that can be applied to the other semesters. There is a sharp decline in `wb.ratio` diagonally across the grids.

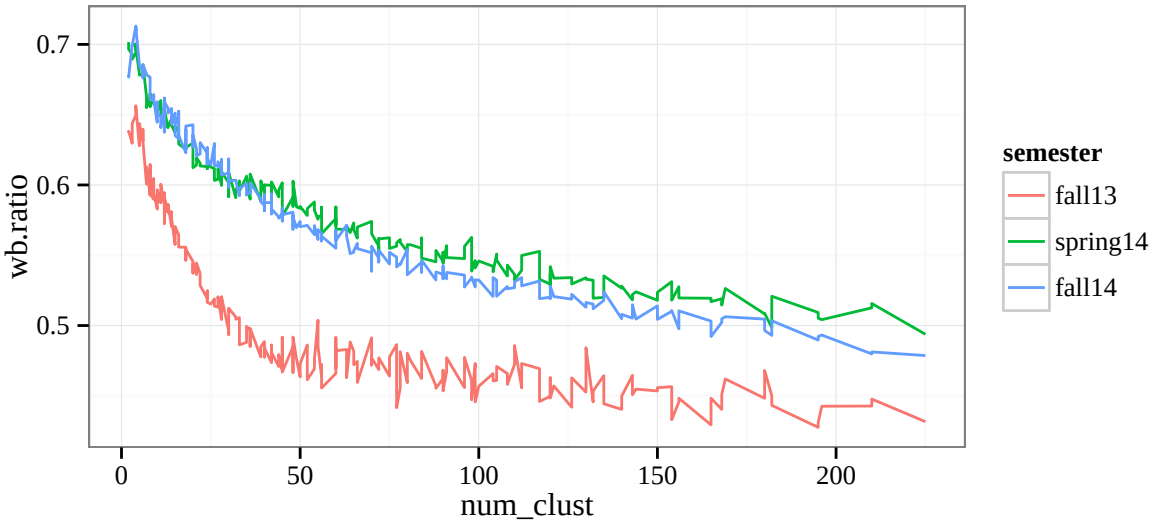

Figure 15: A plot of `wb.ratio` by total number of nodes in the self organizing maps that were fit to homework grades. We focus on fall 2013 in order to create a map that can be applied to the other semesters. There is a sharp decline in `wb.ratio` around 36 nodes in fall 2014 before the ratio levels off. As such, we will work with the 6*imes*6 SOM.

After looking at the `wb.ratio` across the grids and across the total number of nodes, paying special attention to fall 2013, we notice a sharp decline in the ratio at around 36 nodes, or a 6×6 gridded SOM model. Thus, we choose the 6×6 gridded SOM model and map spring 2014 and fall 2014 to our fitted map. However, before mapping our future semesters to the SOM, we first looked at our clustering in the fall semester (figure 17) using a facetted parallel coordinates plot as well as checked convergence (figure 16) looking at the mean distance to the closest SOM unit during training.

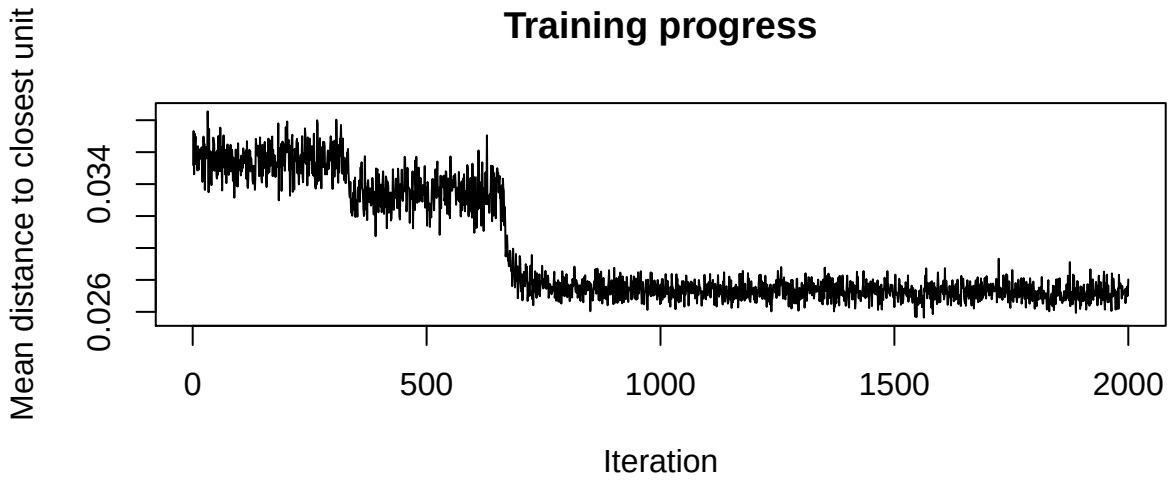

Figure 16: Convergence of our choses 6×6 SOM model over 2000 iterations, looking at the mean distance to the closest SOM unit during training. It appears we have a steady convergence at around 600 iterations.

Looking at the mean distance to the closest SOM unit during training. It appears we have a steady convergence at around 600 iterations, which means we have run our chosen SOM model enough iterations during fitting.

We can see some definite patterns within the clusters, specifically in cluster 31, we have the very good students across the board and within clusters 4 and 5 a steady decline in homework grades past a certain point. Of course, this is not a perfect model by any means as there are definitely some clusters that do not do well, for example cluster 34 has a serious anomoly, and clusters 12 and 18 seem uite varied. Additionally, an issue with this model is that there are 36 clusters, which is a very high number, expecially if our goal is to interpret these clusters in terms of student types.

Now we can fit our other semesters data, spring 2014 and fall 2014, to the chosen fall 2013 SOM model and see if we have similar patterns in the grades. But first, we must process the data to be in the same form as fall 2014, i.e. number of columns and column names. To accomplish this, we use the mapping given in the problem assignment and use simple averages for scores from topics that multiply map to fall 2013 topics.

- Chapter 1, Topic01: What is statistics?
- Chapter 2, Topic02, Topic04: Displaying and Describing Categorical Data
- Chapter 3, Topic03: Displaying and Summarizing Quantitative Data
- Chapter 4, Topic05: Understanding and Comparing Distributions
- Chapter 5, Topic06: The Standard Deviation as a Ruler and The Normal Model
- Chapter 6, Topic07: Scatterplots, Association, Correlation
- Chapter 7,8, Topic08: Linear Regression, Regression Wisdom
- Chapter 9, Topic11: Understanding Randomness
- Chapter 10, Topic09: Sample Surveys
- Chapter 11, Topic10: Experiments
- Chapter 12, Topic12: From Randomness to Probability
- Chapter 15I, Topic15: Sampling Distribution Models - Proportions
- Chapter 16, Topic16: Confidence Intervals for Proportions
- Chapter 17, Topic17: Testing Hypotheses about Proportions
- Chapter 15II, Topic18: Sampling Distribution Models - Means
- Chapter 18, Topic19: Inferences About Means
- Chapter 19, Topic20: More About Tests and Intervals
- Chapter 20, Topic21, Topic22, Topic23: Comparing Groups

Additionally, Chapter 12: From Randomness to Probability was missing from both spring 2014 and fall 2014

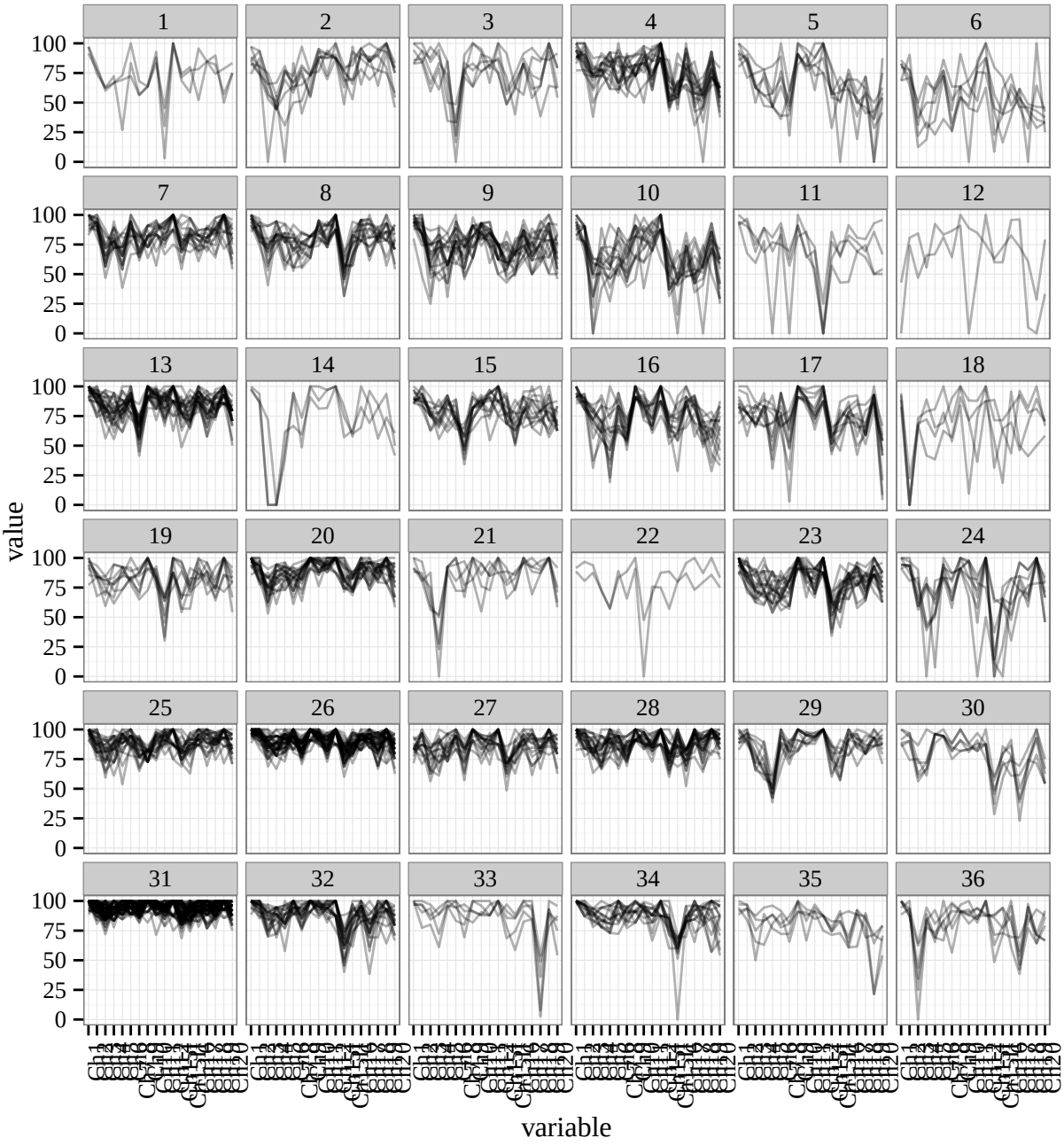

Figure 17: Clustering in the fall semester of our SOM model using a facetted parallel coordinates plot. We can see some definite patterns within the clusters, specifically in cluster 31, we have the very good students across the board and wihtin clusters 4 and 5 a steady decline in homework grades past a certain point.

and Chapter 18: Inferences About Means and Chapter 19: More about Tests and Intervals were both missing from fall 2014. Thus, we threw these variables out of the 6×6 SOM model and refit before prediction. The results of predicting our new data can be seen in figure 18.

From figure 18, we do see some clusters that remain intact in the new semesters, for example cluster 7 appears to be those students that performed very poorly on Chapter 5: The Standard Deviation as a Ruler and The Normal Model. Additionally, clusters 5, 6, and 9 seem to have decent patterns within them. However, overall there is a lot of variation within the clusters, showing that this may not be a very good model for prediction. Additionally, we can look at the `wb.ratio` for this new data, the value of 0.623, which is higher than the `wb.ratios` for fall 2013. This poor prediction leads us to try and perform model based clustering.

Model based clustering

Acknowledgements

We would like to thank Dr. Cook for her advice on dealing with missing values and her imputation code. Also, we used the R packages amap (Lucas 2014), cluster (Maechler et al. 2014), DMwR (Torgo 2010), fpc (Hennig 2014), gdata (Warnes et al. 2014), ggplot2 (Wickham 2009), gridExtra (Auguie 2012), reshape2 (Wickham 2007), and vegan (Oksanen et al. 2015).

References

Auguie, Baptiste. 2012. *GridExtra: Functions in Grid Graphics*. http://CRAN.R-project.org/package=gridExtra.

Hennig, Christian. 2014. *Fpc: Flexible Procedures for Clustering*. http://CRAN.R-project.org/package=fpc.

Iowa State University. 2015. "Blackboard Learn." https://bb.its.iastate.edu.

Lucas, Antoine. 2014. *Amap: Another Multidimensional Analysis Package*. http://CRAN.R-project.org/package=amap.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2014. *Cluster: Cluster Analysis Basics and Extensions*.

Oksanen, Jari, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, and Helene Wagner. 2015. *Vegan: Community Ecology Package*. http://CRAN.R-project.org/package=vegan.

Torgo, L. 2010. *Data Mining with R, Learning with Case Studies*. Chapman; Hall/CRC. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Warnes, Gregory R., Ben Bolker, Gregor Gorjanc, Gabor Grothendieck, Ales Korosec, Thomas Lumley, Don MacQueen, Arni Magnusson, Jim Rogers, and others. 2014. *Gdata: Various R Programming Tools for Data Manipulation*. http://CRAN.R-project.org/package=gdata.

Wehrens, R., and L.M.C. Buydens. 2007. "Self- and Super-Organising Maps in R: The Kohonen Package." *J. Stat. Softw.* 21 (5). http://www.jstatsoft.org/v21/i05.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer New York. http://had.co.nz/ggplot2/book.

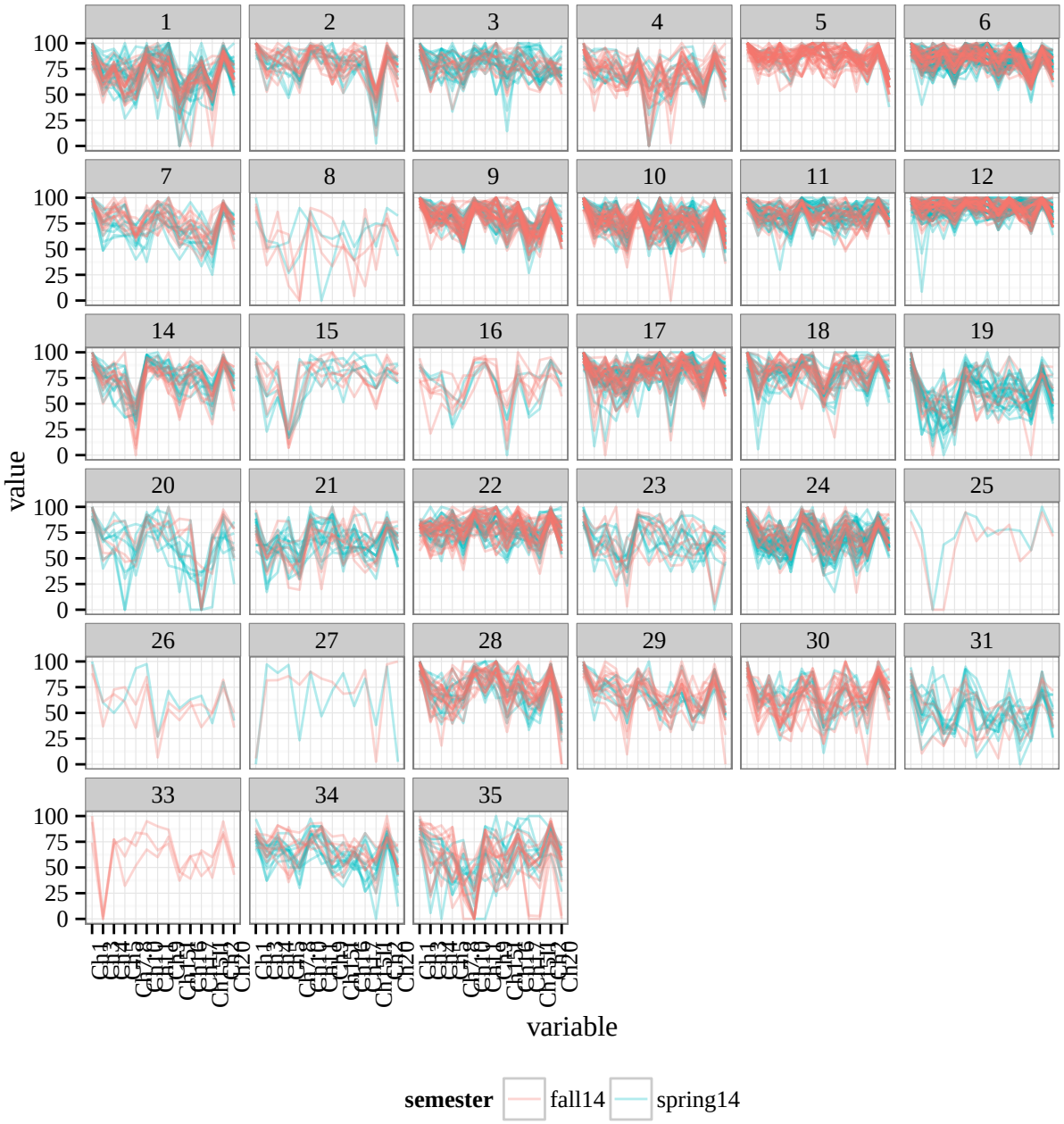

Figure 18: Prediction of spring 2014 and fall 2014 clusters based on the equivalent mapping of homework topics by mapping the data onto the (modified) trained 6×6 SOM model. Nodes 34 and 35 no longer have any members in the cluster for this new data.