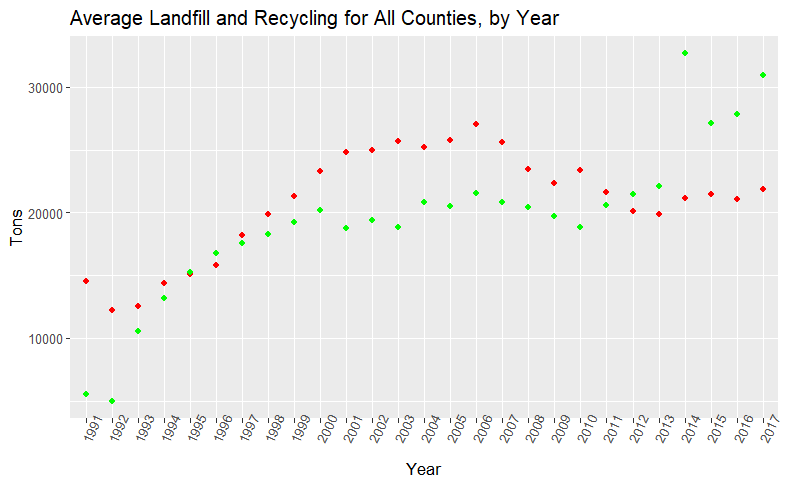
**ANOVA Analysis of Landfill and Recycling Data in Minnesota Counties**

Introduction

This paper examines changes in the amount of recycling and landfill in Minnesota between years 1991 and 2017. Data, grouped by county, is provided by the State of Minnesota Pollution Control Agency. The original dataset provides tonnage for each county, for each year between 1991 and 2017, for both recycling and landfill. For preliminary investigation, graphs were made to observe both recycling and landfill tonnage produced by various groupings. Figure 1 shows the mean value, for all counties combined, of both recycling and landfill, for each year between 1991 and 2017.



Landfill

Recycling

Figure Mean value, for all counties combined, of recycling and landfill for each year between 1991 and 2017.

From the above graph it appears that the amount of recycling increased more than landfill in the last seven years. For this reason, I was interested in whether the ratio of recycling to landfill changed over time.

In particular, the first question this paper addresses is:

* Is the mean value of the ratio of recycling/(recycling +landfilled) tonnage different between the years 1994, 2004, and 2017? If so, which years are significantly different?

The numeric dependent variable, *ratio*, is calculated using the provided data as:

The independent variable for this question is *year,* as a factor, where the three levels are 1994, 2005, and 2017. The ratio values are calculated for each county in Minnesota, for each of the three years of interest.

To answer the question of whether the mean value of the ratio is different between the three years, we have the following hypothesis:

The second question the paper addresses is:

* Is there a difference in the mean value of the ratio, as defined above, for rural versus urban counties? Also, is there an interaction between county type (rural or urban) and year. For this question we will also find the main effect of year, but that will have already been answered in question 1.

For the second question the dependent numeric variable is *ratio*, and the two independent variables are *year* and *countyType*. Year is a factor with three levels: 1994, 2005, and 2017. CountyType is a factor with two levels: urban and rural. As in question 1, the individual ratio values are calculated for each county for each of the three years. For question 2, we have the following hypothesis to test the main effect of countyType:

As mentioned above, for question 2 we could also test the main effect of year, but that has already been done in question 1. The hypothesis for the interaction is:

Methods

For question 1 the single independent variable *year* is repeated three times for each individual county. Since there is only one independent variable, and it is repeated, the appropriate analysis for this type of question is one-way repeated measures ANOVA. After performing ANOVA, a Tukey comparison is done to examine where differences lie.

Chart

Description automatically generatedChart, histogram

Description automatically generatedThe assumption of independence is satisfied as the counties are independent from each other. The assumptions of normality and equal variance are satisfied by the large amount of data. We can see from figure 2 that the dependent variable is reasonably normally distributed for the aggregate of counties and years. Figure 3 shows somewhat normal data for ratio for the individual years.

Figure 3 DV “ratio” for each of the 3 years in the analysis

Figure 2 Histogram of DV “ratio” for all counties and all years

The second question is an extension of the first: In addition to comparing the means of the ratio variable for three years, we are also interested in whether there is a difference in mean ratio based on the type of county (urban or rural). We are also interested in whether there is an interaction between the variables *year* and *countyType*. Since we now have two independent variables, we need a two-way ANOVA. Also, since year is a repeated measure and countyType is fixed, this analysis requires a mixed repeated measure ANOVA.

To determine the designation of rural or urban, we use the Landfilled variable as a proxy for countyType. Figure 4 shows the mean landfill value of all years, for each county. There is a steep decline in landfill tonnage after the three largest counties, so only those counties are considered “urban.” Those counties are Hennepin, Dakota, and Ramsey.

Chart

Description automatically generated

Figure 4 Average landfill of years 1991 – 2017 for every county in Minnesota. The three counties with the most landfill tonnage are Hennepin, Dakota, and Ramsey. Those are designated “urban,” and all other counties are “rural.”

Chart

Description automatically generated with medium confidenceChart, histogram

Description automatically generatedThe assumptions of the analysis for question 2 are met as in question 1: The counties are independent and all ratios are normally distributed in aggregate with equal variance. Looking at the counties separated into rural and urban (figure 5) shows those groups having the DV normally distributed for all years. However, as shown in figure 6, since we are only considering three counties to be urban, when we only consider 3 years it is hard to say that data is normally distributed.

Figure 5 DV “ratios” for urban and rural factors, all years

Figure 6 DV “ratios” for urban and rural, three years

A small amount of data wrangling was necessary to perform the analysis. First, I eliminated all rows where landfilled = 0. The reason for the eliimination was that I did not have confidence in those entries, as it is unlikely any county ever had 0 landfill. There was enough other data so that excluding those rows should not alter the results.

Next, if a county did not have data for one of the three years of interest, that county was eliminated. I took the intersection of all counties with data for 1994, 2005, and 2017. The dataset for question 2 retained 75 of 85 original counties.

Conclusion

Table 1 lists the F statistics and p-values for question 1 and question 2.

|  |  |  |
| --- | --- | --- |
| **F Statistics and P – Values for Question 1 and 2** | | |
|  | **F Statistic** | **P - value** |
| **Question 1** |  |  |
| *Year* | 8.176 | 0.0004 |
| **Question 2** |  |  |
| *Year* | 8.176 | 0.0004 |
| *countyType* | 0.427 | 0.516 |
| *Year : countyType* | 0.977 | 0.379 |

Table F statistics and P – values for both question 1 and question 2

For question 1 we find a p-value of 0.0004. This indicates that at a significance level of 0.05 we should reject the null hypothesis. There is significant evidence to support the claim that at least two mean ratios of recycling/(recycling +landfilled) tonnage are different between the years 1994, 2004, and 2017.

To find where the differences lie a Tukey comparison was done. Table 2 lists the Z statistics and P-values for the comparison between each pair of the three years.

|  |  |  |
| --- | --- | --- |
| **Z Statistics and P – Values for Question 1, Tukey Comparison** | | |
| **2 factors compared** | **Z Statistic** | **P - value** |
| **2005 – 1994** | -0.481 | 0.880 |
| **2017 - 1994** | 3.236 | 0.00353 |
| **2017 - 2005** | 3.718 | < 0.001 |

Table Z statistics and P – values for the Tukey comparison of means.

From the results we see that there is a difference in the mean ratio between years 2017 and 1994, where the p-value is 0.00353. Also, from a p-value of < 0.001, we also see there is a difference in the mean value of ratio for the years 2017 and 2005. There is no significant difference in the mean ratio for the years 2005 and 1994, where the p-value is 0.880.

The results of question 1 are not surprising in view of figure 7, which shows the mean value of the ratio variable for all counties combined, for each year between 1991 and 2017. From the graph we see that the mean value of ratio in 1994 and 2005 were both between 0.4 and 0.42, whereas the value in 2017 was above 0.50. The interesting thing about this graph is the way it rises from 1991 to 1996, then dips down and rises again after 2005.

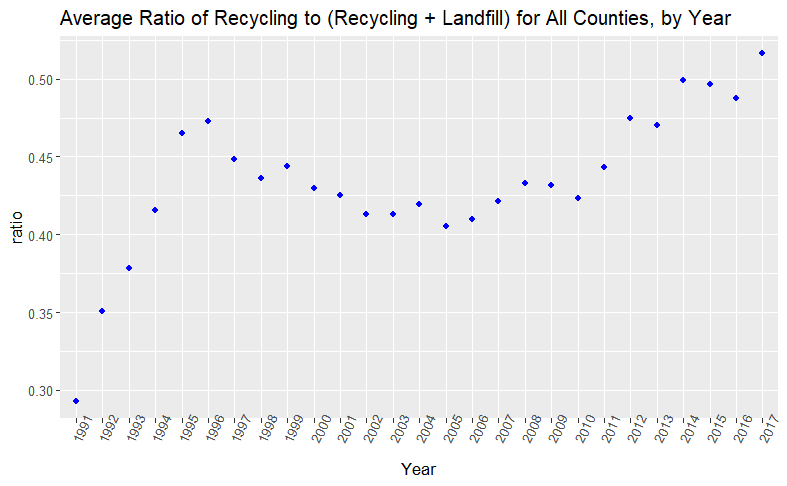


Figure 7 Mean of ratio variable shown for all counties combined, for each year in the dataset.

The dip in the graph of figure 7 can be traced back to figure 1. There we see the value of landfill increasing in 1996 and decreasing from 2006 to 2012. The recycling values trended upward throughout the entire time period. The dip in figure 7 is roughly a mirror image of the landfill in figure 1. This makes sense as landfill is part of the denominator of the ratio variable.

The conclusion for question 2 is that there is no evidence to support the hypothesis that there is a significant difference in mean ratio between urban and rural counties. From table 1 we see that the p-value for the main effect of countyType is 0.516. We already knew from question 1 that the main effect of year was significant, and this shows up again in the results of question 2, where the p-value is again listed as 0.0004. We also find for question 2 that the interaction between countyType and year was not significant, with a p-value of 0.379.

The most interesting thing I found was the result of question 2. The results surprised me somewhat due to the following graph. Figure 8 shows the mean value of ratio, grouped by countyType, for all years. There are trends that exist, and they are different between countyType. I suspect that if I had done the analysis where only years 2005 and 2017 were used, instead of including 1994, we would have seen an interaction between year and countyType, and the main effect of countyType may have been significant.

Chart, scatter chart

Description automatically generated

Figure 8 Mean value of ratio for combined counties, grouped by countyType, for all years

The one limitation of concern for me was only using three counties for the urban class. To claim a normal distribution and equal variance is, I think, limited in this case.