REVISIONS IN UNITED NATIONS ENERGY STATISTICS DATA: CAN CHANGES TO THE PAST IMPROVE OUR UNDERSTANDING OF THE PRESENT?

by

Connor D. Briggs

Honors Thesis

Appalachian State University

Submitted to the Department of Mathematical Sciences
in partial fulfillment of the requirements for the degree of

Bachelor of Science

May, 2021

Approved by:

Eric S. Marland, Ph.D., Chair, Department of Mathematical Sciences, Thesis Director

---

Dennis Gilfillan, Ph.D., Thesis Director

Gregg Marland, Ph.D., Second Reader

William J. Cook, Ph.D., Honors Director, Department of Mathematical Sciences

Abstract

Before data can be analyzed effectively, it needs to be evaluated to determine its quality. With the rise in interest in combating global climate change by reducing carbon dioxide (CO_2) emissions, understanding data quality and certainty is important for producing accurate estimates of emissions from fossil fuel combustion and other industrial processes. These estimates are essential for monitoring the prospects and progress towards national targets for reduction. Accurate fossil fuel CO_2 ($FFCO_2$) estimates, along with climate modeling and current atmospheric conditions, are also essential to forecast future global temperatures. Most fossil CO_2 emissions inventories rely on energy statistics to generate country and year specific estimates of CO_2. Every year, the United Nations publishes an updated Energy Statistics Database that provides annual statistics on the production, processing, trade, and use of fuels for over 230 countries, going back as far as 1950. Each subsequent publication of the database provides additional entries and updated values to the previous years' data. This opens up questions about these data revisions: what changes are taking place, how are countries revising their data, and are there recognizable patterns that provide information to anticipate changes in future database releases? The purpose of this thesis is to learn from the revisions to address questions and concerns surrounding this Energy Statistics Database. This begins with an investigation into the changes occurring in the datasets year-to-year and ends with an inference into patterns within these changes. The additional insight this presents allows a deeper level of understanding for where in the database revisions are occurring, to better characterize the uncertainty of present CO2 value estimates.

Contents

1 INTRODUCTION

The need for reliable data is important when looking at global carbon dioxide (CO_2) estimates. Changes in global temperatures are largely driven by CO_2 emissions, and one of the largest source of anthropogenic CO_2 is the combustion of fossil fuels. One mitigation strategy used to monitor current fossil fuel CO_2 ($FFCO_2$) emissions and subsequent reductions is national level targets for reduction. Monitoring the progress of individual national targets is essential for tracking current prospects and trends of reductions, with the goal of mitigating global climate change. Individual countries track the consumption of fossil fuels to help estimate $FFCO_2$ emissions. However, uncertainty within $FFCO_2$ estimates can be difficult to calculate [2] but is necessary to accurately track national targets in reduction. Because of this, an understanding of the uncertainty within $FFCO_2$ estimates is required.

One framework for examining uncertainty within these estimates is through examining revisions within previous data for fossil fuel usage. The reevaluation of past data can result in revisions needing to be made to past entries. An example of this is in the evolving nature of the United Nations Energy Statistics Database. This database catalogs fossil fuel usage in their original units and these fossil fuel usage values serves as a predecessor for calculating $FFCO_2$ emission estimates [4]. As a result, the reliability of the fossil fuel usage data reported in this database directly impacts the uncertainty of CO_2 estimates calculated from it. The quality of the data being used to calculate CO_2 emissions is an indicator of uncertainty within CO_2 estimates. As such, the revisions within the United Nations Energy Statistics Database present information on where uncertainty may be observable in the data.

The revisions within the United Nations Energy Statistics Database are approached from several directions and methodologies. First, changes made to the reporting and classification of several different fuel commodities are investigated. Beginning in 2011, the United Nations adopted the International Recommendations for Energy Statistics (IRES) to help standardized energy statistics across all countries [7]. This caused the need for administrative changes to occur to data within the database, separate from revisions, to meet the new standardization of energy statistics. Administrative changes made to the classification codes in 2011 are investigated to determine what data inside the database were impacted as a result. Then, revisions are examined from the perspectives of database release to database release, age of the data, fuel group (i.e. solid, liquid, or gaseous fuels), and country the data originated from. Both the frequency and magnitude of the revisions are presented. Finally, the countries within the database are grouped together by the revisions they make using k-means clustering to attempt to determine if certain countries follow closely to each other with revisions, or if they are acting completely uniquely. This will hopefully provide insights into the revisions being made to energy statistics and provide the framework to produce more reliable estimates with an improved understanding of uncertainty.

2 BACKGROUND

2.1 DESCRIPTION OF THE UNITED NATIONS ENERGY STATISTICS DATABASE

The United Nations Energy Statistics Database is maintained by the United Nations Statistics Division (UNSD) and catalogs the production, trade, transformation, and use of solid, liquid, and gaseous fuels, electricity, and heat for over 200 countries. Each entry in the database includes a country where the entry comes from, a commodity code for the type of product, a transaction code for the flow of the product, and the corresponding year of the data. There is then a value attached for the quantity of that entry, with the units dependent on the commodity. Solid and liquid fuels are recorded in thousand metric tons (kt) while gaseous fuels are in terajoules (TJ).

Entries in the database date back as far as 1950. New entries are updated annually as well as revisions made to previous entries. The data are collected by the UNSD via an annual questionnaire submitted by each country. Since the entries in the database are recorded in their original units, the quantity values serve as a predecessor to calculating CO_2 estimates.

In February, 2011, the United Nations adopted the International Recommendations for Energy Statistics (IRES). The aim of this was to, ”provide a comprehensive methodological framework for the collection, compilation and dissemination of energy statistics in all countries irrespective of the level of development of their statistical system.” [7] This came as a result of the work done by the Oslo City Group [5], a working group started in 2006 by the United Nations to further develop energy statistics and standards.

2.2 COMMODITIES AND TRANSACTIONS

The United Nations Energy Statistics Database releases from 2008 through 2018 are currently available for use in this thesis. Within the database a subgroup of commodity and transaction codes were focused on to complete this research. The commodity and transaction codes given in tables 1 through 4 were examined.

Commodity	Code
Hard Coal	CL
Brown Coal	LB
Coke Oven Coke	OK
Brown Coal Briquettes	BB
Patent Fuel	BC
Peat	PT
Gas Coke	GK
Brown Coal Coke	BK
Coal Tar	CT
Other Coal Products	CP
Peat Products	BP
Oil Shale/Oil Sands	OS

Table 1: Solid Fuel Commodity Codes

Commodity	Code
Gas Oil/Diesel Oil	DL
Other kerosene	KR
Motor Gasoline	MO
Bitumen	BT
Conventional Crude Oil	CR
Kerosene-Type Jet Fuel	JF
Liquified Petroleum Gas	LP
Petroleum Coke	PK
Natural Gas Liquids (NGL) n.e.s.	MP
Lubricants	LU
Fuel Oil	RF
Naphtha	NP
Natural Gas Liquids	GL
Other Oil Products n.e.c.	PP
White Spirit and Special Boiling Point Industrial Spirits	WS
Refinery Feedstocks	FS
Paraffin Waxes	PW
Aviation Gasoline	AV
Plant Condensate	CD
Natural Gasoline	NT
Ethane	EA
Gasoline-Type Jet Fuel	GJ
Other Hydrocarbons	OH

Table 2: Liquid Fuel Commodity Codes

Commodity	Code
Natural Gas (Including LNG)	NG
Refinery Gas	RG
Coke Oven Gas	OG
Blast Furnace Gas	BG
Gasworks Gas	GG
Other Recovered Gases	BO

Table 3: Gaseous Fuel Commodity Codes

Commodity	Code
Production	01
Imports	03
Exports	04
Bunkers (International)	05
International Marine Bunkers	051
International Aviation Bunkers	052
Stock Changes	06
Non-Energy Uses	11
Flared and Vented	104

Table 4: Transaction Codes

2.3 NON-ZERO CHANGES

Revisions in this research are considered to be non-zero changes in the quantity of an entry. This includes additions, deletions, and quantity changes. Additions are entries added for past years to a new release of the database and represent the quantity going from zero to a new value. Deletions are entries taken out from previous database releases and represent an entry going from some value to a quantity value of zero. Quantity changes are entries that already existed in a previous database release having their quantity changed to a new value.

When looking at frequency of revisions, the total count of non-zero changes is being used. When looking at magnitude changes for revisions, this is referring to a directional movement in the quantity. Additions are a positive, deletions are a negative, and quantity changes are positive or negative depending on the change in value. New entries added to a database, where the entry year is that of the database year or further forward, are not classified as additions. These are considered new entries for that database release and are therefore not revisions.

2.4 DATABASE YEAR

When referring to the database editions, the year of the database is used. This refers to the year the database release coincides with, not the year the database became available. As of this thesis, the most recently made available database is that for 2018. There is usually a lag of 2-3 years between the collection of data and its reporting in a database. For example, the database published in 2020 includes data through 2018 and is identified here as the 2018 edition.

2.5 HISTORIC AND NEAR REVISIONS

Two ways that revisions to entries are examined throughout this research are by the age of the entry and whether it is a historic or near revision. The age of an entry refers to the difference in years between the database year and the entry's year. New entries to the database with the same year as the database year start with an age of 0. Historic revisions are classified as revisions made to entries whose age is greater than or equal to 10 years. Near revisions are then revisions made to entries less than 10 years old.

For example, in the 2008 database, new entries with the year 2008 would have an age of 0. Revised entries for the year 2007 would have an age of 1, a year of 2006 would have an age of 2, and so on up until 10, which would be the year 1998. Once the age of an entry hits 10 years or more, it is labeled as historic.

3 ADMINISTRATIVE CHANGES TO COMMODITY CODES IN THE 2011 DATABASE

3.1 BACKGROUND TO ADMINISTRATIVE CHANGES

As a result of the implementation of the IRES, the 2011 release of the United Nations Energy Statistics Database brought about multiple changes to the data codes. Several of the commodity codes that appeared prior to the 2011 release were discontinued for use. The discontinued commodity codes this thesis focused on include CD, MP, and NT for liquid fuels and BK for solid fuels.

3.2 REMOVED LIQUID FUEL CODES

Commodities CD, MP, and NT are secondary liquid fuels. Their codes correspond to plant condensate, natural gas liquids (NGL) n.e.s., and natural gasoline, respectively. The entries for the imports, exports, stock changes, and non-energy uses of these commodities appear to have been moved to the commodity code GL, natural gas liquids, in the 2011 release. This was determined by matching the entries for CD, MP, and NT in the 2010 database to entries in the 2011 database. Entries with a quantity value equal to zero were also not used since they had a high chance of finding false positive matches.

In total, the 2010 database contained 651 entries for CD, 1552 entries for MP, and 1,239 entries for NT that had non-zero quantities, for a total of 3,442 entries. These entries were matched against entries in the 2011 database that did not appear in the 2010 database, excluding new entries added for 2011. This was done through the “setdiff” function from the dplyr package, which returns all the rows in the 2011 database that do not have identical matches in the 2010 database. The entries in the 2010 database were then matched by their country, transaction, year, and quantity using the “match_df” function from the plyr package in R to this differenced data from the 2011 database. In total, tables 5 through 7 give a breakdown of the matches found by transaction.

Transaction	CD Frequency (2010)	CD Matches (2011)
Production (01)	415	0
Imports (03)	24	18
Exports (04)	128	92
Stock Changes (06)	51	19
Non-Energy Uses (11)	33	33

Table 5: Frequency of CD Entries by Transaction

Transaction	MP Frequency (2010)	MP Matches (2011)
Production (01)	762	15
Imports (03)	189	175
Exports (04)	336	253
Stock Changes (06)	202	197
Non-Energy Uses (11)	63	62

Table 6: Frequency of MP Entries by Transaction

Transaction	NT Frequency (2010)	NT Matches (2011)
Production (01)	969	0
Imports (03)	13	7
Exports (04)	158	129
Stock Changes (06)	99	67
Non-Energy Uses (11)	0	0

Table 7: Frequency of NT Entries by Transaction

As can be seen, a large percentage of the entries for these removed commodity codes were able to be found for all transaction types, except for production. It is unclear where the entries for these three for production were displaced. No substantial evidence for their existence within the 2011 release could be determined and it is possible they were removed altogether. Of the matches found for imports, exports, stock changes, and non-energy uses, the breakdown of the matched commodities is given in table 8.

Commodity	CD Matches (2011)	MP Matches (2011)	NT Matches (2011)
CL	0	1	0
CR	0	8	0
GJ	0	1	0
GL	162	692	203

Table 8: Commodity Matches for Removed Liquid Fuel Codes

One limitation with using the "match_df" function for this application is the possibility of false positive matches. If multiple entries in the 2011 database had the same country, transaction, year, and quantity, they would be detected as a match. This is why entries with zero quantity values were removed, due to the abundance of observations that would encounter this issue. Detecting matches with this method also has no way of detecting observations that have a quantity change between the 2010 and 2011 releases. Looking at the matches for the MP observations that did not match to a GL observation, it is plausible these are false detections due to their sparsity and irregularity. GL is the commodity code for natural gas liquids, a primary liquid fuel. From these findings it is probable that the observations with commodity codes CD, MP, and NT in the 2010 database were moved to the GL commodity code in the 2011 database, at least for imports, exports, stock changes, and non-energy uses.

Transaction	2010	2011	Difference	Percent Change
Production (01)	4,353,485	574,602	3,778,883	-87%
Imports (03)	637,308	607,722	-29,586	-5%
Exports (04)	1,874,112	1,634,370	-239,742	-13%
Stock Changes (06)	9,798	9,923	125	1%
Non-Energy Uses (11)	121,351	121,509	158	0%

Table 9: Change in Magnitude (kt) for Removed Liquid Fuel Codes

One last way to support this theory is to look at the change in magnitude that occurred across all GL observations (table 9). The total magnitude of the quantities for CD, MP, and NT in the 2010 database were summed up by transaction type. The same was done for the GL observations that were new to the 2011 database, excluding observations for that year. The change in magnitude supports the idea that the entries were moved to the GL commodity for all but production. The total change in magnitude of entries with the transaction code for production was looked at for all other commodities in the 2011 database and none indicated a change that could explain where the production observations went.

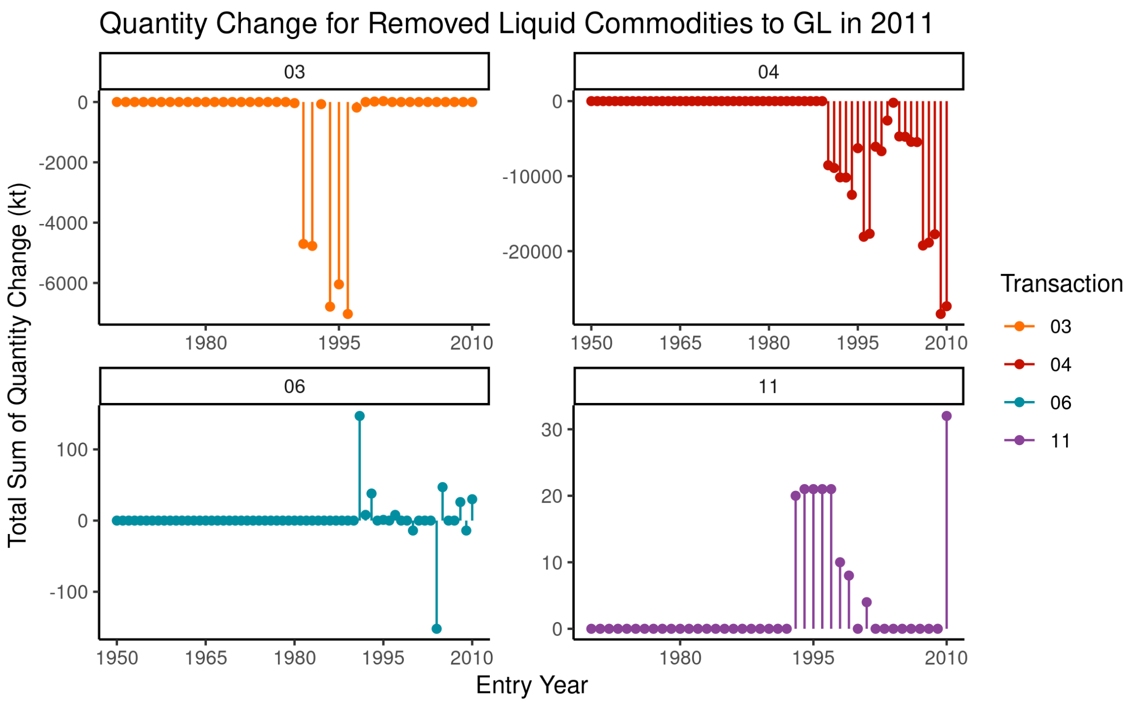

Figure 1: Change in Magnitude from CD, MP, & NT (2010) to GL (2011) by Year

The difference between the quantity sum of revisions in GL entries in 2011 database and the quantity sum of the removed entries from the 2010 database were plotted (figure 1). This looks at the sum of the quantity for new or revised values for GL in the 2011 database by year and subtracts the sum of CD, MP, and NT entries from the 2010 database for that year. If the entries for the removed liquid codes were moved to GL with no other revisions made to data from that year, the value should be at or near zero. Every point is also influenced by any other revisions made in the 2011 database to GL entries for that year, which could explain the change in magnitude for some years.

3.3 REMOVED SOLID FUEL CODES

BK was the commodity code for brown coal coke, a secondary solid fuel. In total, 345 entries with a non-zero quantity value for BK existed in the 2010 database. Following a similar process as with the removed liquid codes resulted in the following matches (table 10).

Transaction	BK Frequency (2010)	BK Matches (2011)
Production	126	3
Imports	76	9
Exports	96	39
Stock Changes	42	3
Non-Energy Uses	8	1

Table 10: Frequency of BK Entries by Transaction

While a much smaller percentage of matches were found than with liquid fuels, all 55 of the matches belonged to the commodity code OK, coke oven coke, a secondary solid fuel. It is likely based off of this that OK is the new residency for the BK observations.

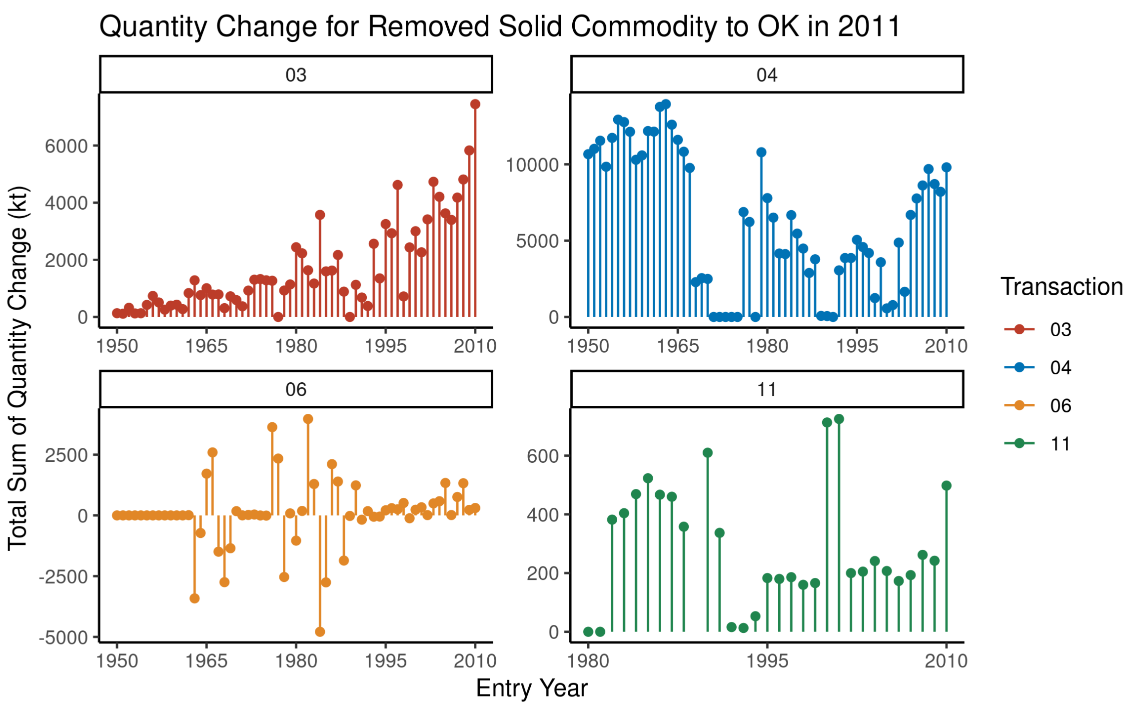

Figure 2: Change in Magnitude from BK (2010) to OK (2011) by Year

The change in magnitude for the OK entries in the 2011 database also shows some correlation to the removal of the BK entries (figure 2). The sum of the quantity for each year of OK entries in the 2011 database not in the 2010 database was found. The sum of the quantity for each year of BK entries from the 2010 database was also found and subtracted from this. Points at zero for each year would indicate the BK entries were moved to OK. However, other revisions to OK entries will also be reflected in this.

4 NON-ZERO CHANGES

4.1 THE EFFECT OF AGE ON THE DATABASE AND ITS ENTRIES

Revisions within the database take place with each new edition and any entries within the database are susceptible to revisions. Even entries in the database dating back to 1950 are subject to revision. Because of this, when comparing any of the past databases to a more current release, the disparity between the two will only increase with age. A new edition of the database will never look more in line with a prior one, due to the evolving nature of the data.

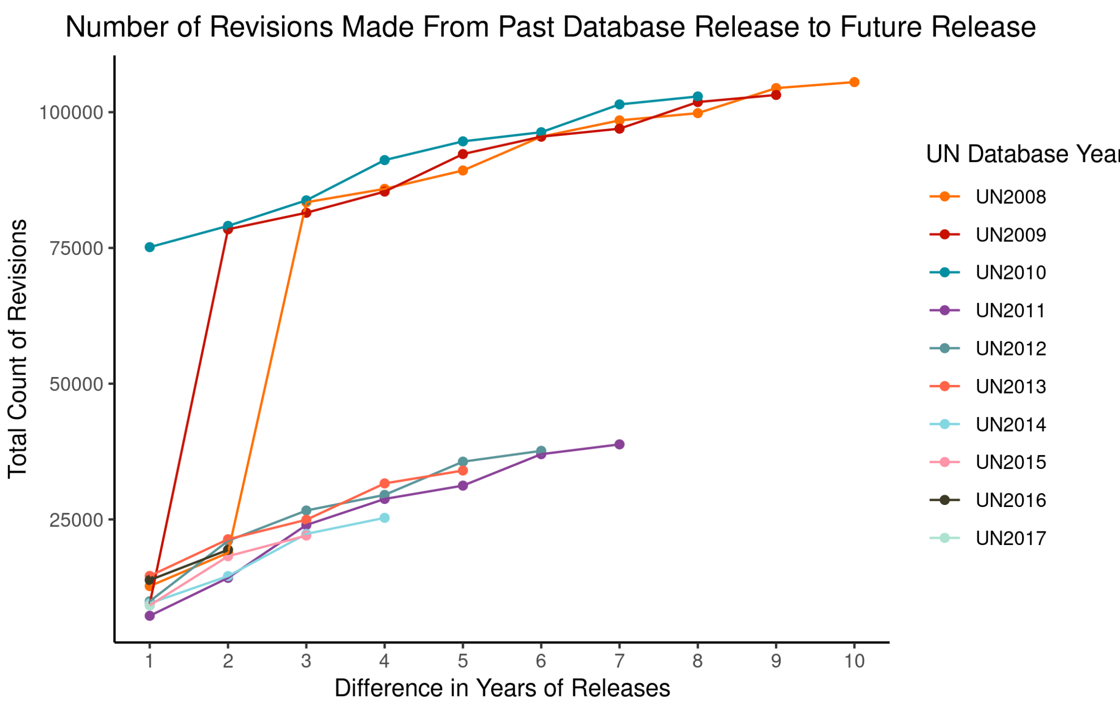

Figure 3: Frequency of Revisions After Database Release

The graph in figure 3 demonstrates the changing nature of a database relative to all the databases after. Each database, starting with 2008, was individually compared to each release after that year. The axis for age is the difference in years from the database's year to the age of the database being compared. The axis for count is the total count of revisions made from the starting database to the ending. This includes additions, deletions, and quantity changes; but does not include new entries for later releases into the count.

The large increase in revisions for the 2008, 2009, and 2010 databases is a result of the administrative changes that occurred in 2011. Due to the methodology of the comparison, each subsequent database will have one less database to compare to than the database before it. The plot clearly shows a positive correlation between the age and number of revisions that have occurred, for both databases before and after the changes in 2011. The database demonstrates that revisions will steadily and continually be made after the initial release.

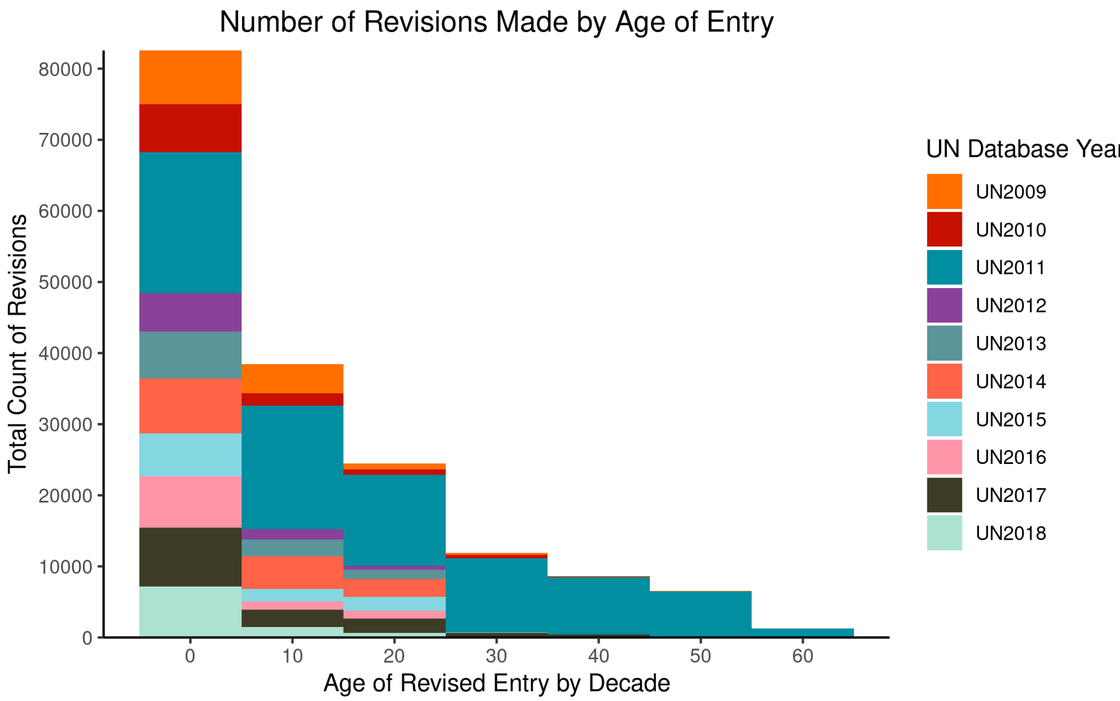

Figure 4: Frequency of Revisions in Each Database by Entry Ages

Looking at the age of entries being revised is also an important indicator as to what is happening within the database. The histogram in figure 4 visualizes the number of revisions happening within each ten-year gap of age in the databases. The age of each revision relative to the database year it originates from was computed and the frequency of the age of revisions was summed up for each ten-year gap of time. Each bar in the plot represents ten years. The first bar corresponds with near revisions, and the second through sixth bars are then historic revisions. The frequency of revisions decreases substantially with each decade older the entries are relative to the database year. The frequency of revisions is fairly consistent between databases for the three bars, representing data from age 0 through 29 years. Aside from the 2011 database, the frequency of revisions to entries past the third bar substantially diminishes.

Near revisions (leftmost bar, figure 4) are made far more frequently in the database than historic revisions (all other bars, figure 4). When looking at the frequency in which near and historic revisions were made (figure 5), it can be observed that, on average, more near revisions are made and that there is a much larger range in the frequency of historic revisions made.

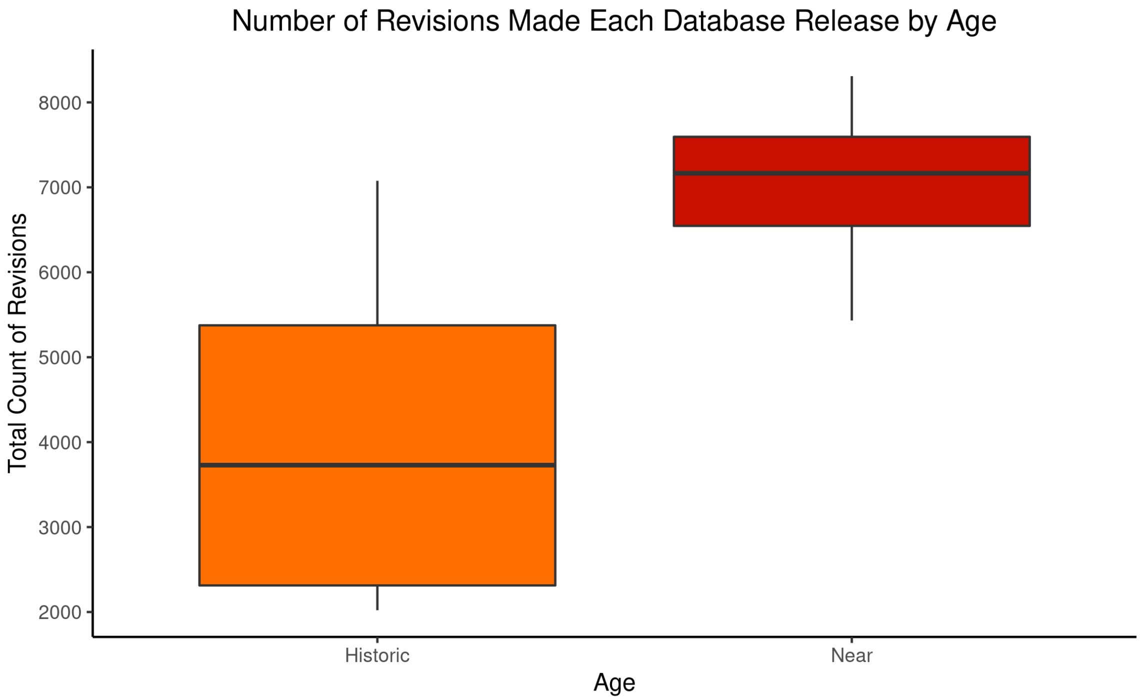

Figure 5: Frequency of Revisions in Each Database by Age

The age of both the database and the age of entries relative to their database play an important role in forecasting the number of revisions. The older a database release is, the less likely it is to resemble a newer release, due to revisions within the database causing it to constantly move further away from older releases. The change year over year when comparing one database to all the subsequent releases is quite consistent over time. Meanwhile, the frequency of revisions being made relative to the age of the entry has a negative relationship. Fewer revisions seem to occur to entries that are older relative to the database year, but they can still occur. A substantial amount of the revisions within the database occur to entries that are less than a decade old, with this number quickly diminishing.

4.2 CHANGES WITHIN FUEL GROUPS

The revisions occurring within the United Nations Energy Statistics Database are not only dependent on the age of entries, but also on the commodity. Commodities are split into solid fuels, liquid fuels, and gaseous fuels (tables 1 through 3). These fuel groups each have their own characteristics and trends when it comes to revisions by frequency and magnitude.

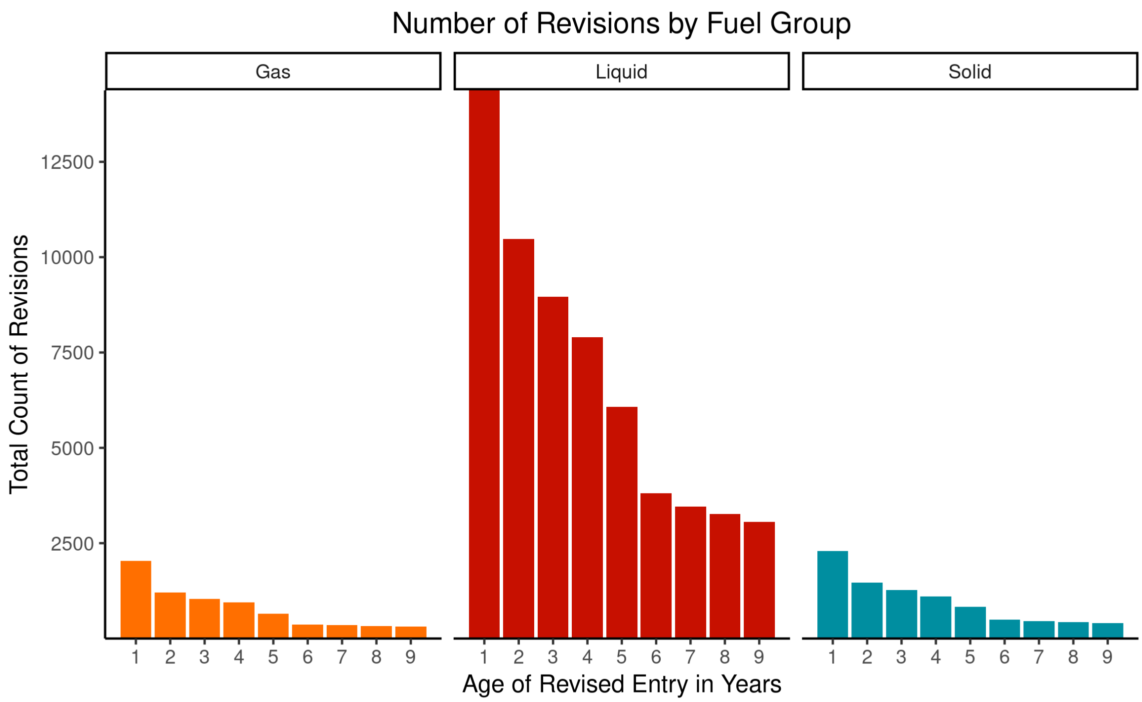

Figure 6: Frequency of Year-to-Year Revisions by Age for Each Fuel Group

Observing the frequency of revisions in each fuel group reveals that there is a negative relationship between the age and frequency of revision of entries. The bar plot in figure 6 shows the total frequency of near revisions in all the databases. The bars are the frequency of all year-to-year revisions made across all editions of the database for entries of a given age. Liquids overwhelmingly have the largest frequency of revisions relative to solids and gasses. All three fuel groups show a sharp drop off with age before beginning to level off toward the end.

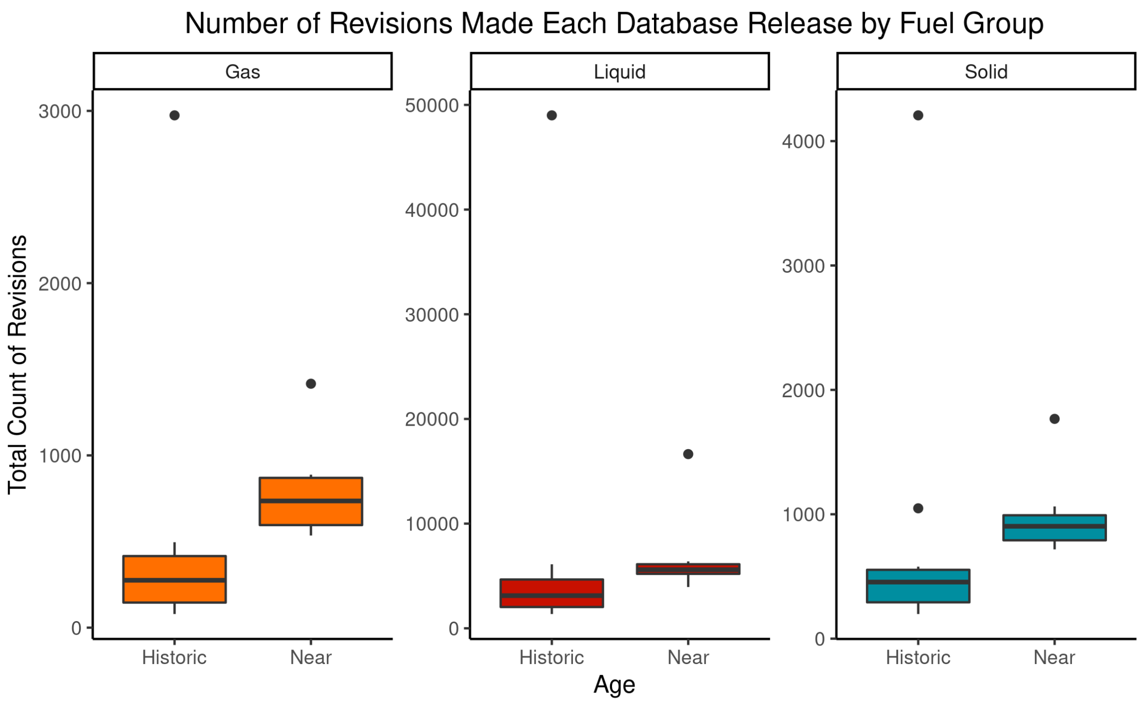

Figure 7: Frequency of Year-to-Year Revisions in Each Database by Fuel Group

Comparing the frequency between near and historic revisions reveals that the databases typically have more near than historic revisions for all fuel groups (figure 7). The boxes are the frequency of revisions made in each database release. In each of the fuel groups, the 2011 database is an outlier, with the frequency of historic revisions in 2011 being much higher than that of near revisions. The range in the number of revisions seen in table 11 for historic revisions is also much larger for all fuel groups. While on average there are more near revisions, the variation in frequency of historic revisions is much greater.

Fuel Group	Age	Minimum Frequency	Mean Frequency	Maximum Frequency
Gas	Historic	79	528	2,974
	Near	535	783	1,416
Liquid	Historic	1,363	7,771	49,000
	Near	3,944	6,507	16,641
Solid	Historic	198	831	4,207
	Near	718	968	1,767

Table 11: Frequency of Year-to-Year Revisions in Each Database by Fuel Group

Fuels within the United Nations Energy Statistics Database are reported in different units, depending on the commodity. Because of this, direct comparisons between the quantities of commodities are not always possible to do in their original units. However, within the commodities focused on in this thesis, the units are consistent within solid, liquid, and gaseous fuels. This allows comparisons of magnitude within each fuel group. Magnitude is the directional change made to the sum of the quantity values. The magnitudes reported here are the sum of the magnitude of change for all revisions within a fuel group. The sign of change is included so that a positive change in one value would be negated by a negative change in another. Adding to the quantity values through additions or increased value changes will yield a positive magnitude, whereas deletions or value changes that decrease the quantity will contribute negatively to magnitude.

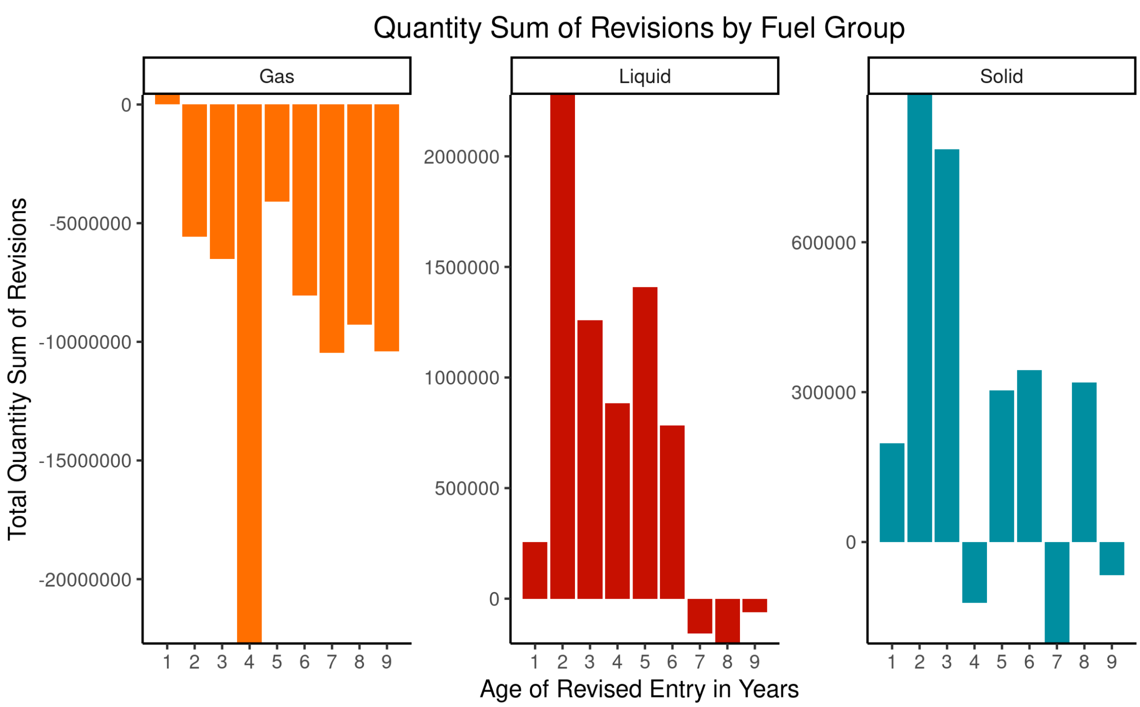

Figure 8: Total Magnitude of Year-to-Year Revisions by Age for Each Fuel Group

When looking at the magnitudes of revisions for near revisions (figure 8), some interesting discoveries can be made. The bars in this plot represent the sum of the magnitude changes for all the database releases. Gaseous fuels are in terajoules (TJ), and solid and liquid fuels are in thousand metric tons (kt). Revisions to gaseous fuels have a generally negative direction, whereas liquids are generally positive for the first few years. Revisions to solid fuels do not seem as consistent, with the magnitude mostly being positive over time.

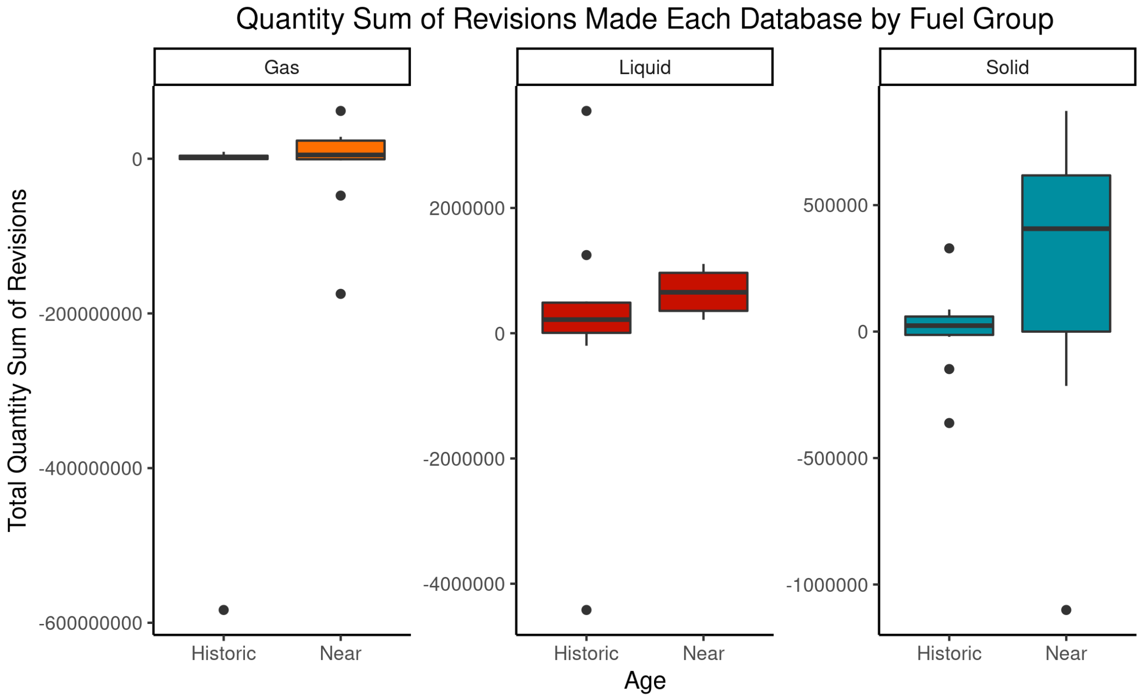

Figure 9: Total Magnitude of Year-to-Year Revisions in Each Database by Fuel Group

The change in magnitude between historic and near revisions for fuel groups (figure 9) is not as conclusive as frequency. Magnitude changes in historic data have many outliers present. Liquids and solids have outliers distributed above and below, while gasses have one outlier of an extreme magnitude in the negative direction. The average magnitude of revisions for gasses is negative, liquids positive, and solids positive for historic and near revisions (table 12). The range for historic revisions is larger for gasses and liquids, but the inverse is true with solids.

Fuel Group	Age	Minimum Magnitude	Mean Magnitude	Maximum Magnitude
Gas (TJ)	Historic	-583,458,430	-55,851,507	8,998,951
	Near	-174,708,382	-7,667,596	61,841,709
Liquid (kt)	Historic	-4,419,181	161,781	3,549,399
	Near	215,353	644,731	1,106,673
Solid (kt)	Historic	-361,573	3,723	329,053
	Near	-1,100,662	245,362	872,348

Table 12: Total Magnitude of Year-to-Year Revisions in Each Database by Fuel Group

The fuel group plays an interesting role in revisions. Across all fuel groups, near revisions are more frequent than historic. The only exception to this is the 2011 database where a massive number of historic entries were changed. With near revisions, the majority of revisions happen within the first few years and then the frequency levels out. There is a lot more variability in the magnitude of revisions. The overall directional trend for gasses is negative, and liquids and solids positive, but this can vary wildly with each database edition.

4.3 REVISIONS IN CONSECUTIVE DATABASE RELEASES BY COUNTRY

The reporting of entries for the United Nations Energy Statistics Database occurs at the country level. As such, the variation between countries when it comes to year-to-year revisions is significant. The top ten countries by frequency for historic and near revisions (tables 13 & 14) reveal that developed countries are typically the ones doing the most revising to entries. The overlap between the top ten countries by frequency is large, with the Netherlands, Japan, United States, Canada, and United Kingdom appearing on both.

Rank	Country No.	Country	Frequency
1	528	Netherlands	2,841
2	392	Japan	2,727
3	840	United States	1,629
4	484	Mexico	1,459
5	158	Taiwan	1,436
6	124	Canada	1,391
7	40	Austria	1,288
8	554	New Zealand	1,169
9	826	United Kingdom	1,091
10	372	Ireland	1,064

Table 13: Top Ten Countries by Frequency of Historic Revisions

Rank	Country No.	Country	Frequency
1	124	Canada	2,239
2	528	Netherlands	1,895
3	392	Japan	1,764
4	364	Iran	1,536
5	36	Australia	1,354
6	826	United Kingdom	1,328
7	40	Austria	1,295
8	251	France	1,223
9	840	United States	1,179
10	579	Norway	1,132

Table 14: Top Ten Countries by Frequency of Near Revisions

Even with the large frequency of revisions for these countries, they still only represent a small fraction of the total number of revisions (figures 10 & 11). The top ten revising countries represent less than a fifth of the total number of revisions for historic and near data. The mean of historic and near revisions for a country are 362 and 351, respectively. While these top ten countries are making on average far more revisions than other countries, they do not dominate the landscape for the frequency of revisions taking place.

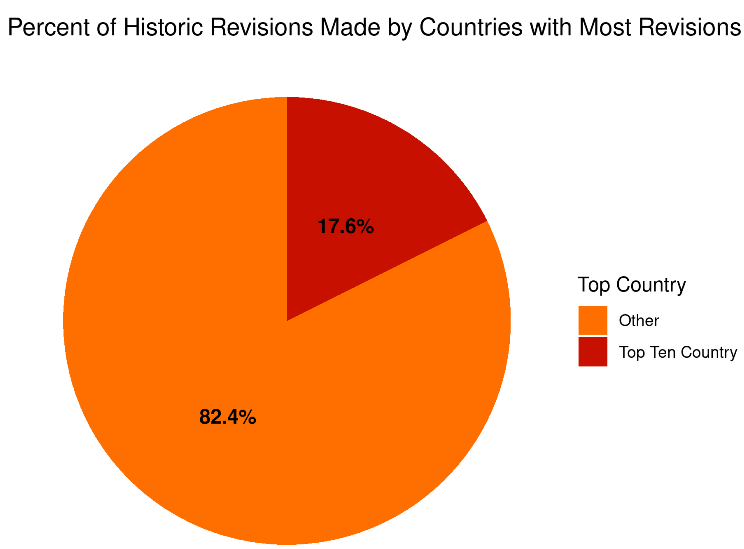

Figure 10: Percent of Historic Revisions by Top Countries

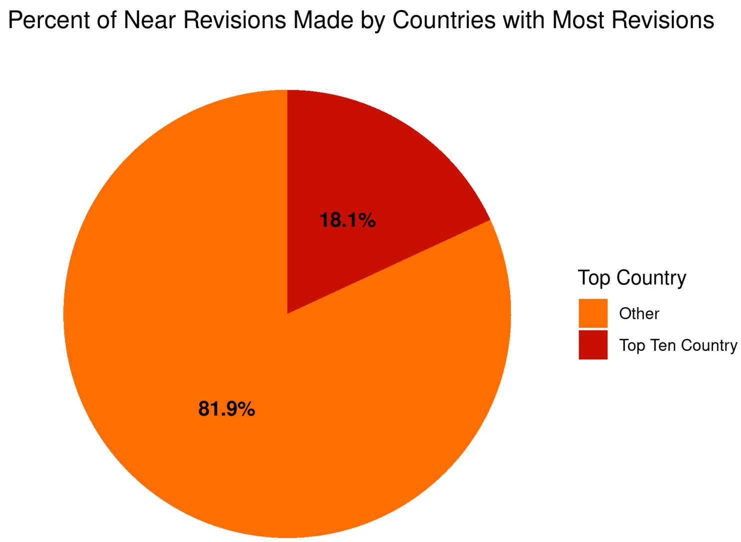

Figure 11: Percent of Near Revisions by Top Countries

The total magnitude of revisions done by country (tables 15 through 20) provides interesting insight into the types of changes individual countries are making. One of the most easily noticeable facts is that Japan is the top country for increases in magnitude from revisions for solids, liquids, and gasses. With Japan being one of the top countries for frequency of revisions, it makes sense that they would most likely also be a top country in magnitude. However, what is surprising is the massive margin by which Japan dominates positive revisions to solids and gasses in comparison to other countries. In those two fuel groups, Japan has almost twice the magnitude as the second ranked country.

Another interesting note is that while China was not a frequent revising country, the magnitude of its revisions placed it in the ranking for solid, liquid, and gaseous fuels. This means that China's average change in magnitude for revisions made must be much larger than that of the countries on the frequency ranking. The same can be said about South Korea and Russia, although South Korea only appears in the top of the solid and liquid fuels ranking in magnitude. Russia only appears in the gaseous fuels. Besides these three countries, all the other countries in the top increases also appeared on the frequency rankings.

Rank	Country No.	Country	Revision Magnitude (kt)
1	392	Japan	1,438,025
2	156	China	731,418
3	410	South Korea	302,498
4	826	United Kingdom	203,289
5	124	Canada	149,087

Table 15: Top Five Countries by Total Magnitude of Revisions to Solid Fuels

Rank	Country No.	Country	Revision Magnitude (kt)
1	392	Japan	2,635,888
2	840	United States	2,145,070
3	410	South Korea	1,012,866
4	156	China	663,475
5	528	Netherlands	661,439

Table 16: Top Five Countries by Total Magnitude of Revisions to Liquid Fuels

Rank	Country No.	Country	Revision Magnitude (TJ)
1	392	Japan	39,995,379
2	528	Netherlands	16,879,344
3	251	France	16,409,004
4	156	China	13,997,007
5	643	Russia	12,016,806

Table 17: Top Five Countries by Total Magnitude of Revisions to Gaseous Fuels

The top countries for decreases in magnitude by fuel group is somewhat unexpected, with very few countries that were on the top of the frequency rankings appearing. This implies that magnitude revisions and frequency revisions are not necessarily driven by the same actors. The Netherlands, Mexico, United States, Iran, and Canada were all on the frequency ranking, but the rest were not. Another interesting point is how for solids, liquids, and gasses the country with the largest magnitude of decreases in revisions for each is around twice that of the second largest. This is a similar trend to what was seen for the largest increases in magnitude.

Rank	Country No.	Country	Revision Magnitude (kt)
1	408	North Korea	-536,336
2	360	Indonesia	-225,314
3	528	Netherlands	-105,362
4	608	Philippines	-68,530
5	643	Russia	-66,544

Table 18: Bottom Five Countries by Magnitude Sum of Revisions to Solid Fuels

Rank	Country No.	Country	Revision Magnitude (kt)
1	12	Algeria	-844,275
2	682	Saudi Arabia	-489,672
3	484	Mexico	-244,120
4	226	Equatorial Guinea	-151,233
5	398	Kazakhstan	-137,379

Table 19: Bottom Five Countries by Magnitude Sum of Revisions to Liquid Fuels

Rank	Country No.	Country	Revision Magnitude (TJ)
1	840	United States	-269,414,153
2	12	Algeria	-104,665,396
3	862	Venezuela	-74,109,507
4	364	Iran	-58,985,573
5	124	Canada	-36,033,491

Table 20: Bottom Five Countries by Magnitude Sum of Revisions to Gaseous Fuels

The revisions being made by countries differ heavily in both frequency and magnitude. Countries that appeared on the frequency rankings, but not the magnitude, could possibly have either made many small changes that did not amount to enough change to be on top, or the changes they made canceled out directionally. In contrast, countries that were on the magnitude ranking but not the frequency were possibly making less frequent but stronger directional revisions, or their fossil fuel usage is orders of magnitude higher than other nations.

5 K-MEANS CLSUTERING OF COUNTRIES

5.1 METHODOLOGY OF K-MEANS CLUSTERING

The countries in the United Nations Energy Statistics Database were grouped together using k-means clustering. The clustering was done for historic and near revisions within each fuel group. Countries were grouped based on their total magnitude of revisions to each commodity within the fuel group and revision age. To do this, all the revisions across the database were split by their commodity into solids, liquids, and gasses, and then by age into historic and near revisions. This resulted in a total of six different datasets of countries and commodities. The magnitude of revisions for every country was then summed up and normalized by commodity. From there, the datasets were ready for k-means clustering.

To determine the groupings of the clusters, the "kmeans" function from the stats package in R was used. This function aims to minimize the sum of squares distance within each cluster, for a specified number of clusters. By default, the function uses the Hartigan and Wong algorithm for clustering. This algorithm determines several random starting points to be the cluster centers and then converges from there to find a combination of centers and groupings that minimized total distance from the points to their respective centers. In each case, twenty-five randomly chosen iterations of centers are chosen. All the other points are assigned to the closest center, and the iteration with the smallest within-cluster sum of squares is then chosen as the starting combination of clusters. Points are selectively reassigned to other clusters until assigning a point to a different cluster would increase the total within-cluster sum of squares [3]. A seed was also set for the clustering, so that the randomly chosen centers were consistent every time it was run.

To determine the number of clusters for each k-means clustering, the within-cluster sum of squares was plotted for each dataset for one through ten clusters. This is the total sum of the squares of the distance from each point to its cluster center. The within-cluster sum of squares will not increase when the number of clusters is increased. As a result, a bend in the plot is looked for to determine a number of clusters where diminishing improvements are obtained by increasing the number further. The number of groupings used for historic and near revisions ended up being the same for each fuel group.

After an optimal number of clusters was determined and the datasets were clustered, the distance matrix for the commodities between every country was determined using the Euclidean distance. Classical multidimensional scaling was used on this matrix to convert it into a 2-dimensional representation by using the ”cmdscale” function from the stats package in R. Since the datasets were on the dimension of the number of commodities within each one, graphical representation was not possible. Classical multidimensional scaling is best when Euclidean distances are used and returns values such that the amount of variation captured in the axis of each commodity gives proportional importance to how much that commodity impacts the scaling. To check how well the scaling represents the dataset, the goodness-of-fit for the scaling can be examined. The goodness-of-fit scales from 0 to 1 and a higher value represents the scaling is doing a better job representing the variance in each dimension.

5.2 CLUSTERING COUNTRIES BY HISTORIC AND NEAR REVISIONS IN SOLID FUELS

The countries in the United Nations Energy Database were clustered by revisions to solid fuel commodities for historic and near revisions. Four clusters were used when clustering both historic and near revisions. The within-cluster sum of squares did not have a large break for either historic or near revision groupings (figures 12 & 13). However, the plots did have a slight diminishing to improvement after four clusters.

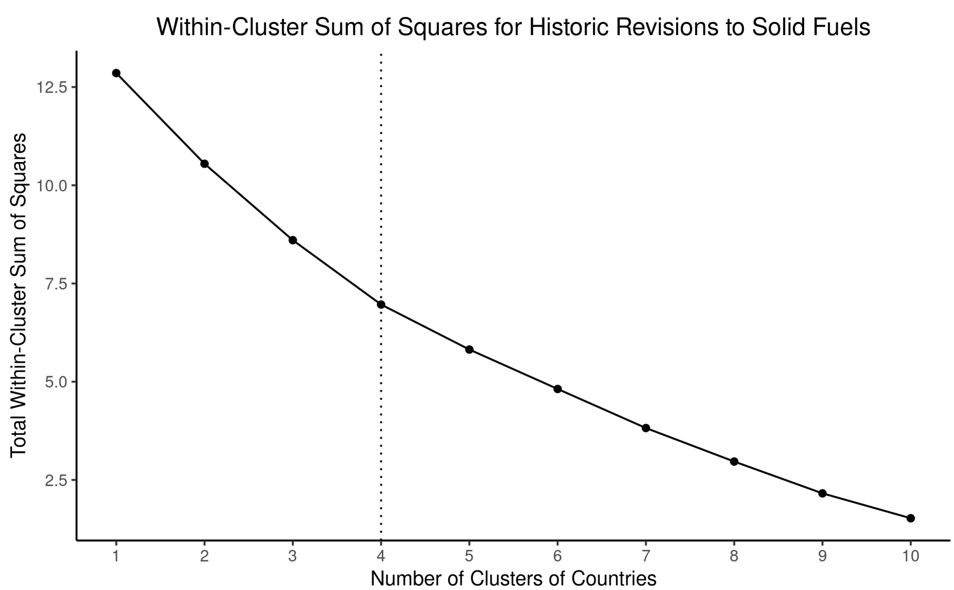

Figure 12: Within-Cluster Sum of Squares for Historic Revisions to Solids

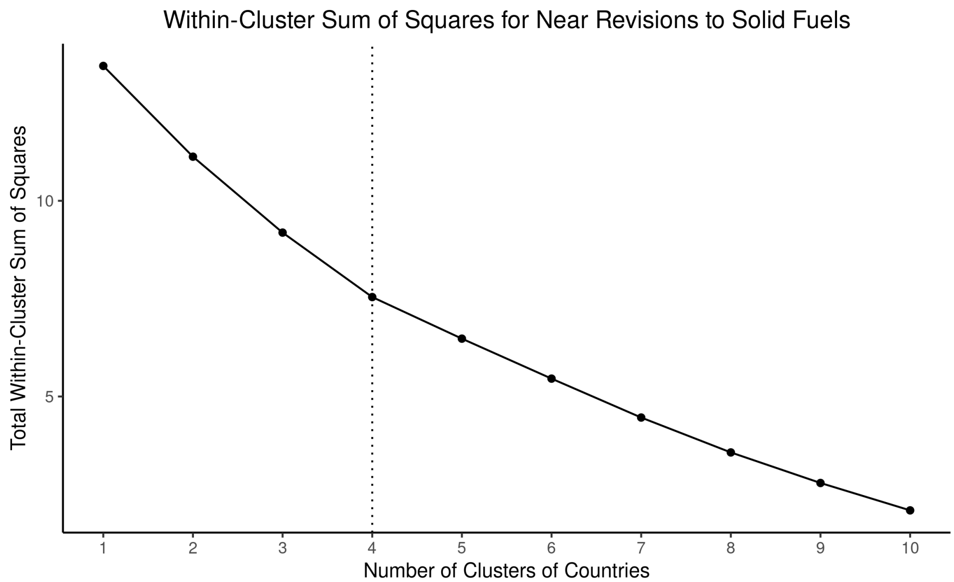

Figure 13: Within-Cluster Sum of Squares for Near Revisions to Solids

For both historic and near k-means clusters (figure 14), a large central cluster was formed, with only a handful of countries that did not fit in being outside of it. The groupings of countries not in the main cluster are consistent between historic and near revisions. In both, the United States (840) and China (156) are their own clusters, and then Belarus (112), Finland (246), and Ireland (372) are clustered together.

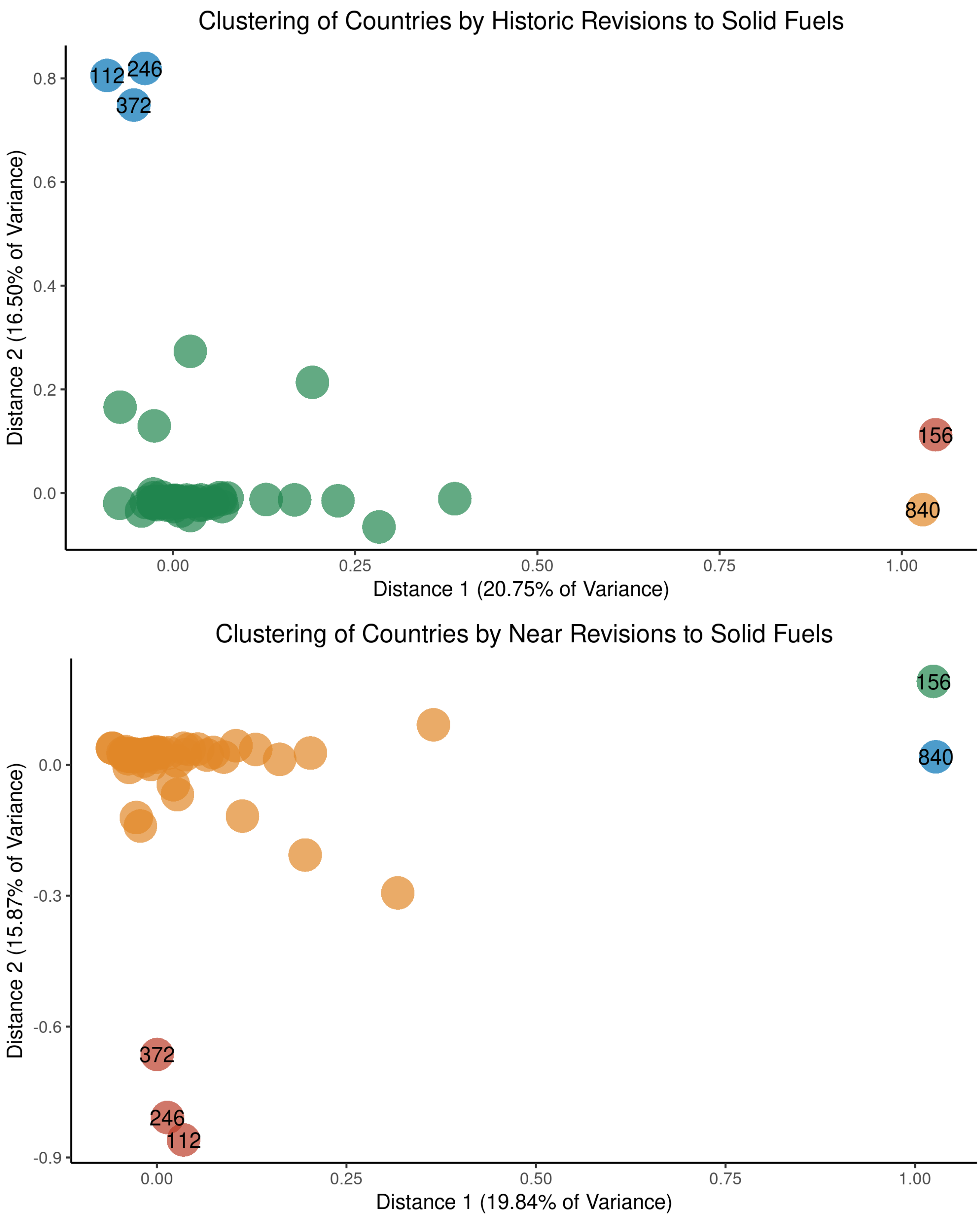

Figure 14: Clustering of Countries by Revisions to Solid Fuels

The goodness-of-fit for the multi-dimensional scaling of the k-means clusters indicates the axes for the historic clusters shows 37.25% of the variance and for the near clusters shows 35.71%. These fits are not good and indicate that the visual representation of the distances between countries are not being properly represented.

5.3 CLUSTERING COUNTRIES BY HISTORIC AND NEAR REVISIONS IN LIQUID FUELS

For liquid fuels, three clusters were used for both k-means groupings. The bend in the plots for within-cluster sum of squares for liquid fuels (figures 15 & 16) was much more pronounced. This made determining the number of clusters an easier decision than with solids.

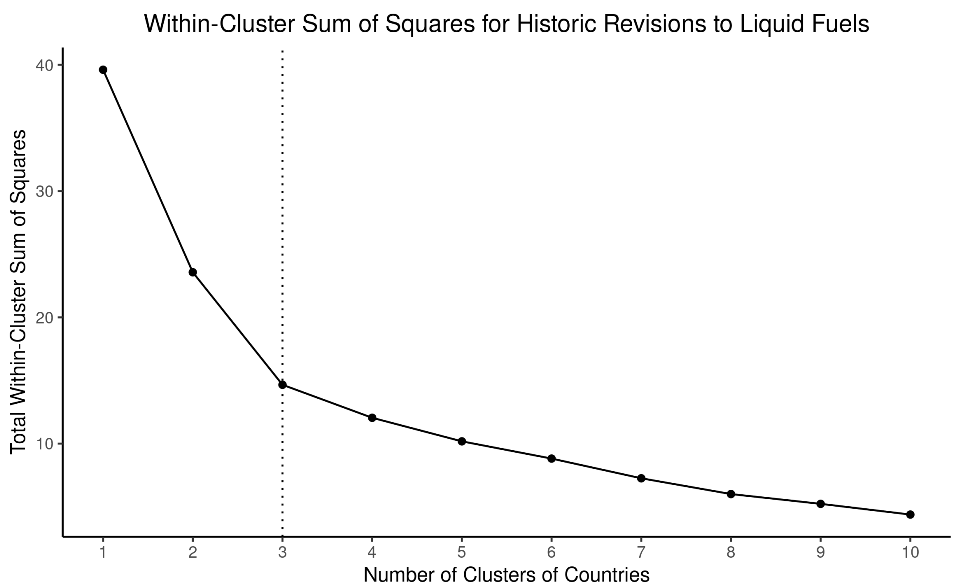

Figure 15: Within-Cluster Sum of Squares for Historic Revisions to Liquids

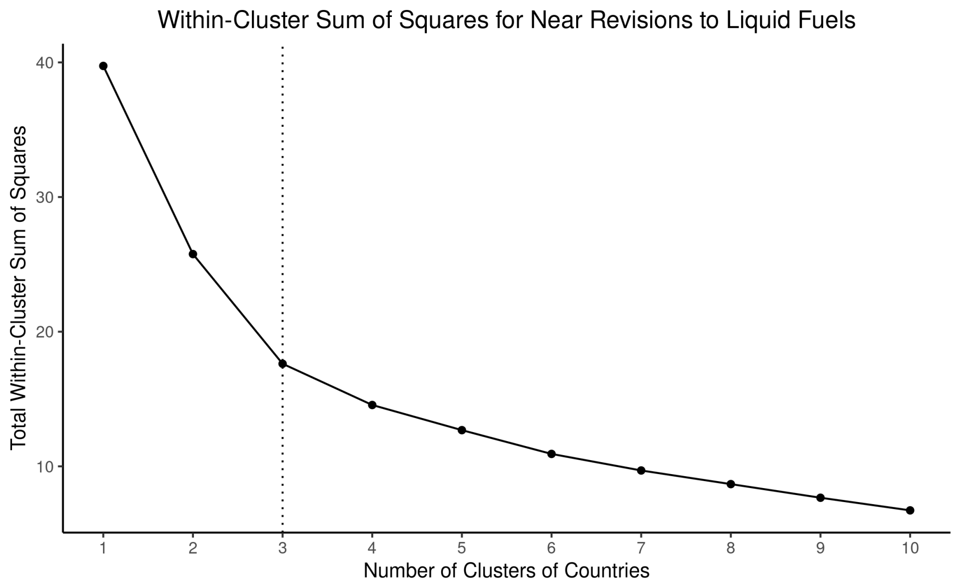

Figure 16: Within-Cluster Sum of Squares for Near Revisions to Liquids

Like with solids, the majority of countries for liquid fuels fell into one large cluster, and then countries with differences falling outside this cluster (figure 17). In both historic and near revisions, the United States (840) was in its own cluster. The other countries outside the main cluster were grouped together. These countries are Algeria (12), Brazil (76), Canada (124), China (156), France (251), Germany (276), India (356), Iran (364), Italy (382), Japan (392), South Korea (410), the Netherlands (528), Russia (643), Saudi Arabia (682), Singapore (702), and the United Kingdom (826) for historic revisions. Near revisions had all the same countries and then additionally Indonesia (360), Kuwait (414), and Mexico (484).

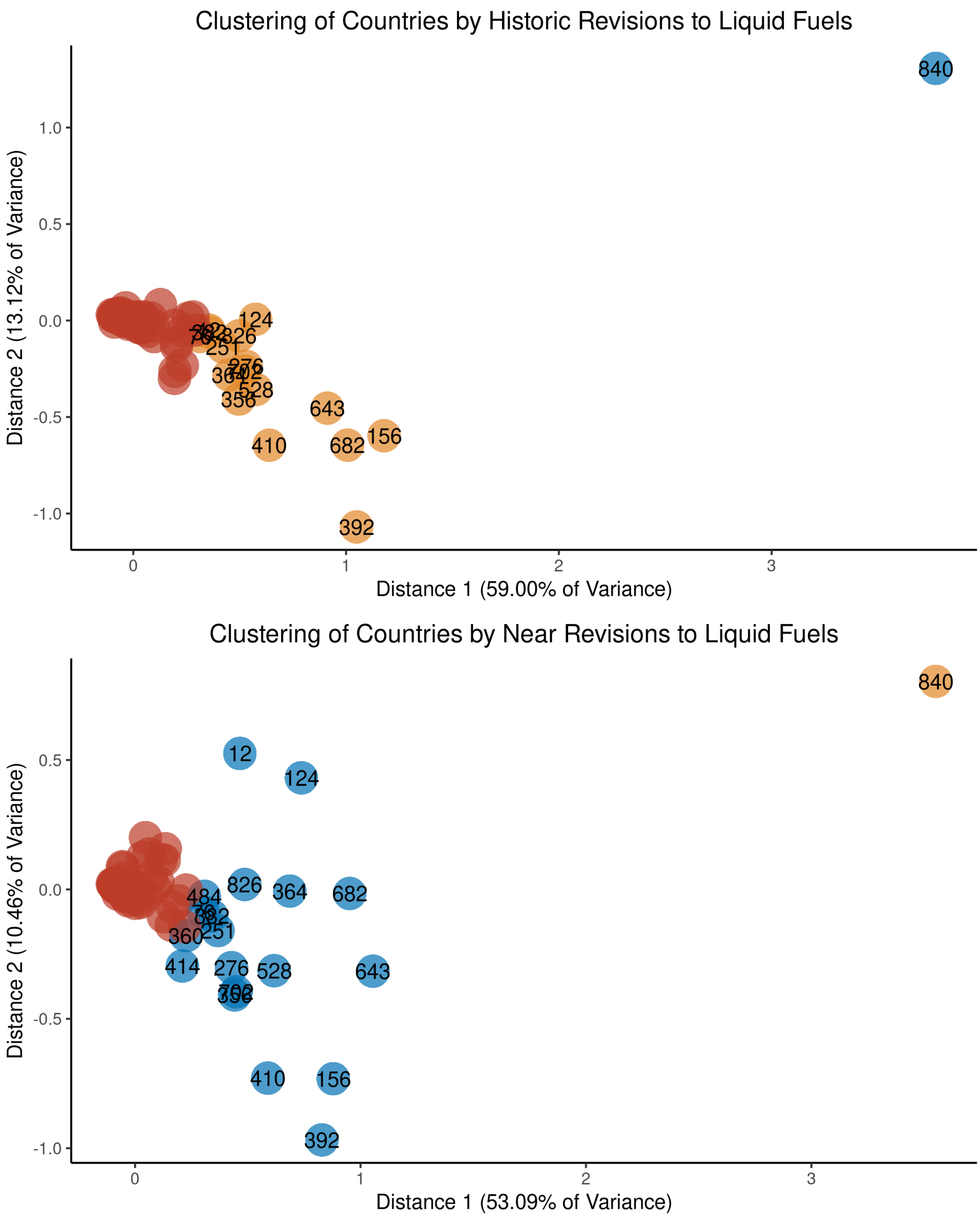

Figure 17: Clustering of Countries by Revisions to Liquid Fuels

The goodness-of-fit values indicate 72.12% of the variance in the historic clusters and 63.55% in the near clusters is shown by the scaling. This is a decent improvement over the solid fuels and means the plots better represent the actual distances between revisions in countries.

5.4 CLUSTERING COUNTRIES BY HISTORIC AND NEAR REVISIONS IN GASEOUS FUELS

Gaseous fuels, like liquids, had a much more prominent bend in the within-cluster sum of squares than solids did (figures 18 & 19). Four clusters were used for both historic and near groupings.

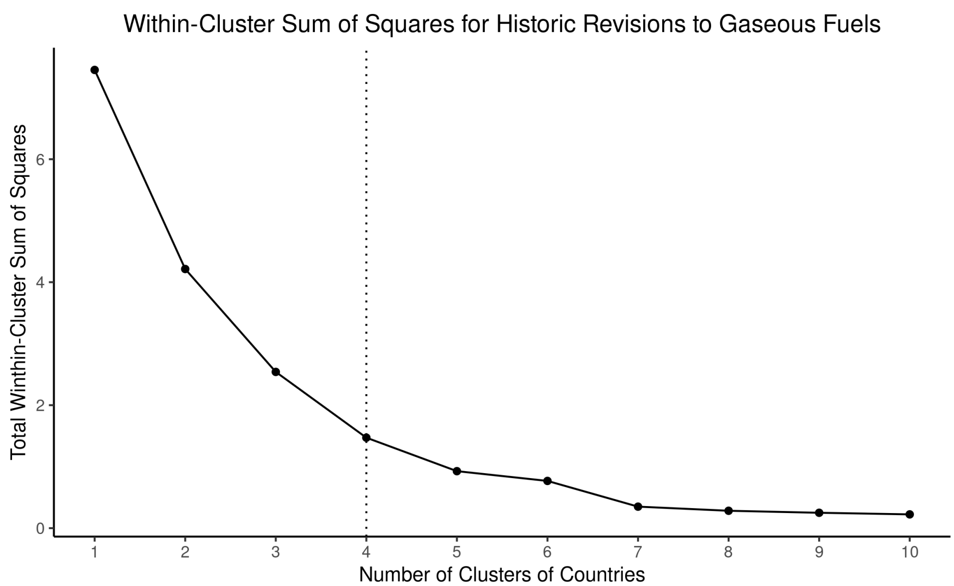

Figure 18: Within-Cluster Sum of Squares for Historic Revisions to Gasses

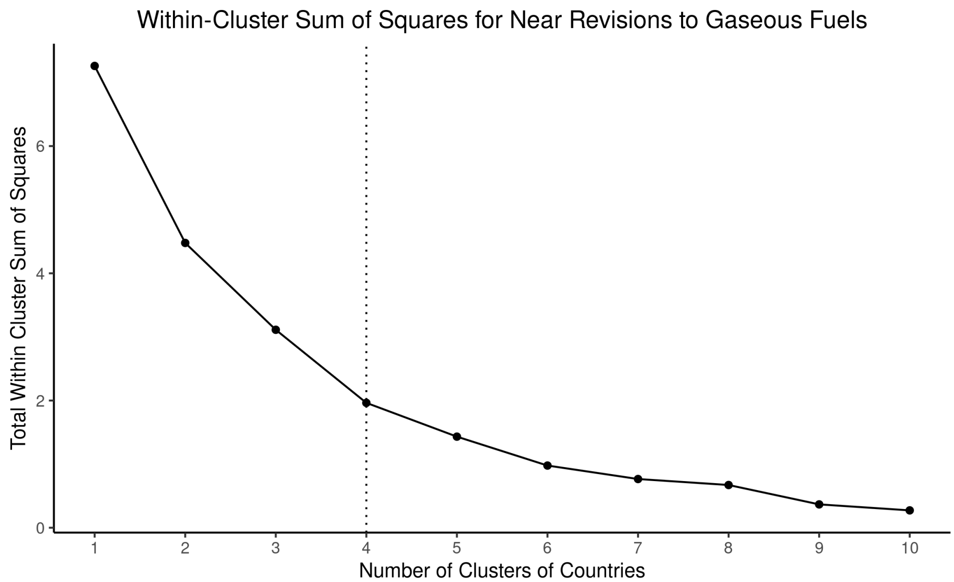

Figure 19: Within-Cluster Sum of Squares for Near Revisions to Gasses

Consistent with what was observed in solids and liquids, the majority of countries when clustering by gaseous revisions fell into the same group (figure 20). This time China (156) was its own cluster, Japan (392) was its own, and Russia (643) and the United States (840) were together for both near and historic revisions.

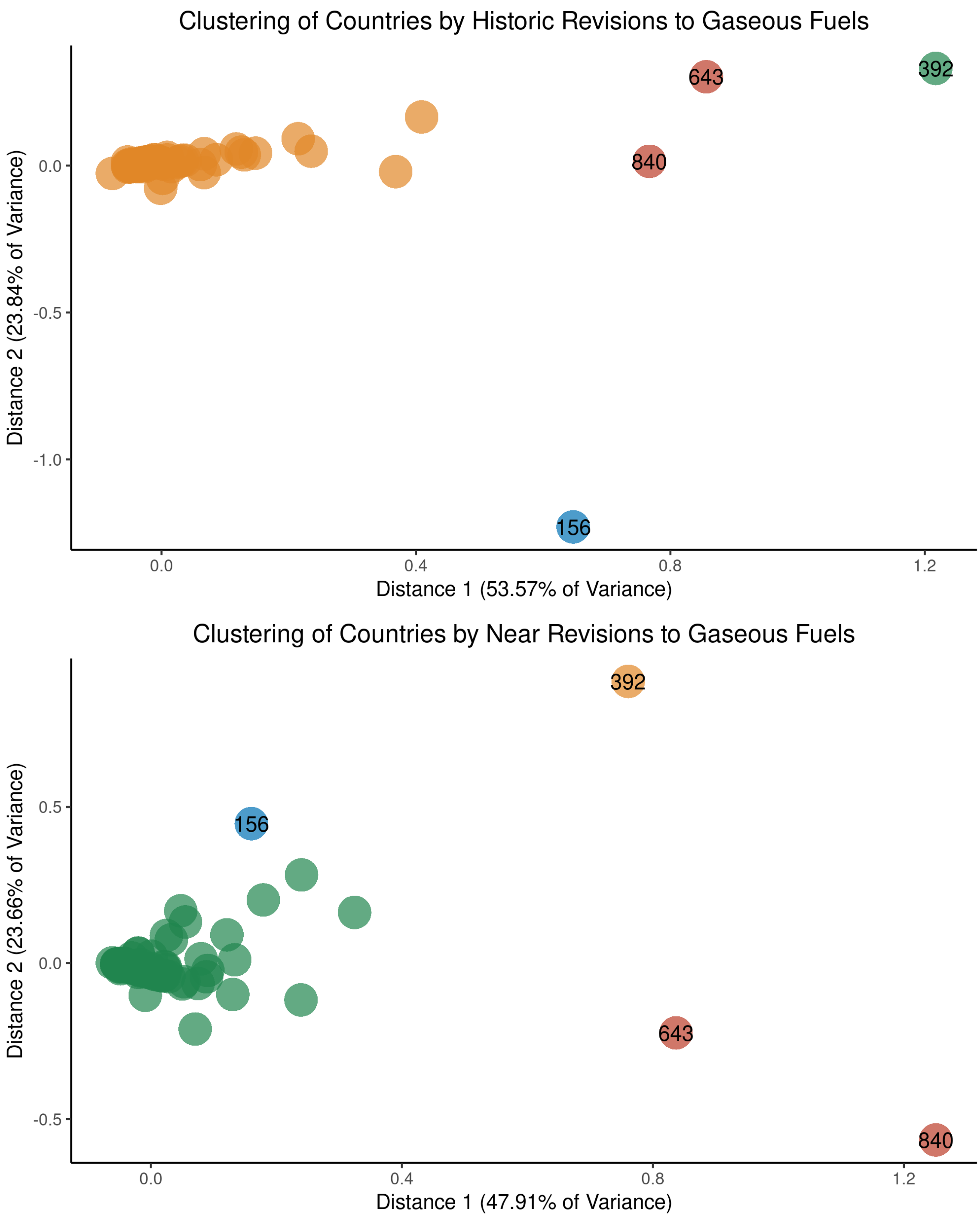

Figure 20: Clustering of Countries by Revisions to Gaseous Fuels

The goodness-of-fit values indicate 77.41% of the variance in the historic clusters and 71.57% of the variance for near clusters is represented by the scaling. These are the best values observed across all the groupings.

5.5 AN EXAMINATION OF COUNTRIES BY CLUSTER

Across all the k-means clusters, almost all of the countries fell into one cluster each time. Each time then, there were only several countries that did not fall into this main cluster. What this implies is that the within-cluster sum of squares was dramatically improved by removing just a handful of countries from the main cluster in each analysis. While increasing the number of clusters would continue to improve the within-cluster sum of squares, these few countries made the bulk of the difference in their respective clustering analysis. The breakdown of countries not in the main cluster is given in tables 21 through 23.

Age	Cluster 1	Cluster 2	Cluster 3
Historic	China (156)	Belarus (112), Finland (246), Ireland (372)	United States (840)
Near	China (156)	Belarus (112), Finland (246), Ireland (372)	United States (840)

Table 21: Countries Not in Largest Cluster for Solid Fuel Revisions

Age	Cluster 1	Cluster 2
Historic	United States (840)	Algeria (12), Brazil (76), Canada (124), China (156), France (251), Germany (276), India (356), Iran (364), Italy (382), Japan (392), South Korea (410), Netherlands (528), Russia (643), Saudi Arabia (682), Singapore (702), United Kingdom (826)
Near	United States (840)	Algeria (12), Brazil (76), Canada (124), China (156), France (251), Germany (276), India (356), Indonesia (360), Iran (364), Italy (382), Japan (392), South Korea (410), Mexico (484), Netherlands (528), Russia (643), Saudi Arabia (682), Singapore (702), United Kingdom (826)

Table 22: Countries Not in Largest Cluster for Liquid Fuel Revisions

Age	Cluster 1	Cluster 2	Cluster 3
Historic	Russia (643), United States (840)	China (156)	Japan (392)
Near	Russia (643), United States (840)	China (156)	Japan (392)

Table 23: Countries Not in Largest Cluster for Gaseous Fuel Revisions

These tables instantly make it clear the United States and China are very different from other countries when it comes to revisions. The United States occupies its own cluster for both of the solid and liquid fuel clusters, and then shares one with Russia for the gaseous fuels. China likewise has its own cluster for both the solid and gaseous fuels, and then is with several other countries in the liquid fuel clustering. Not surprisingly, a large portion of these countries appear in the top of the rankings for revisions by frequency, magnitude, or both.

The countries that were outside the largest cluster of their respective k-means analysis were then investigated further. Boxplots of the total quantity sum of revisions were examined for these countries. It was determined from the k-means analysis that these countries are revising data in a manner that makes them uniquely identifiable from other countries. Because of this, the interest in these countries lies in the commodities in which they are acting differently from what is standard. To look into these scenarios, boxplots in which the countries not within the main cluster are outliers were investigated. This investigation allows commodities in which these countries differ greatly from other countries to be focused on. Outliers were determined using Tukey's method of identifying values outside the range of $[Q1 - 1.5 * IQR,\ Q3 + 1.5 * IQR]$, where IQR is the interquartile range and $Q1$ and $Q3$ are the first and third quartiles [6].

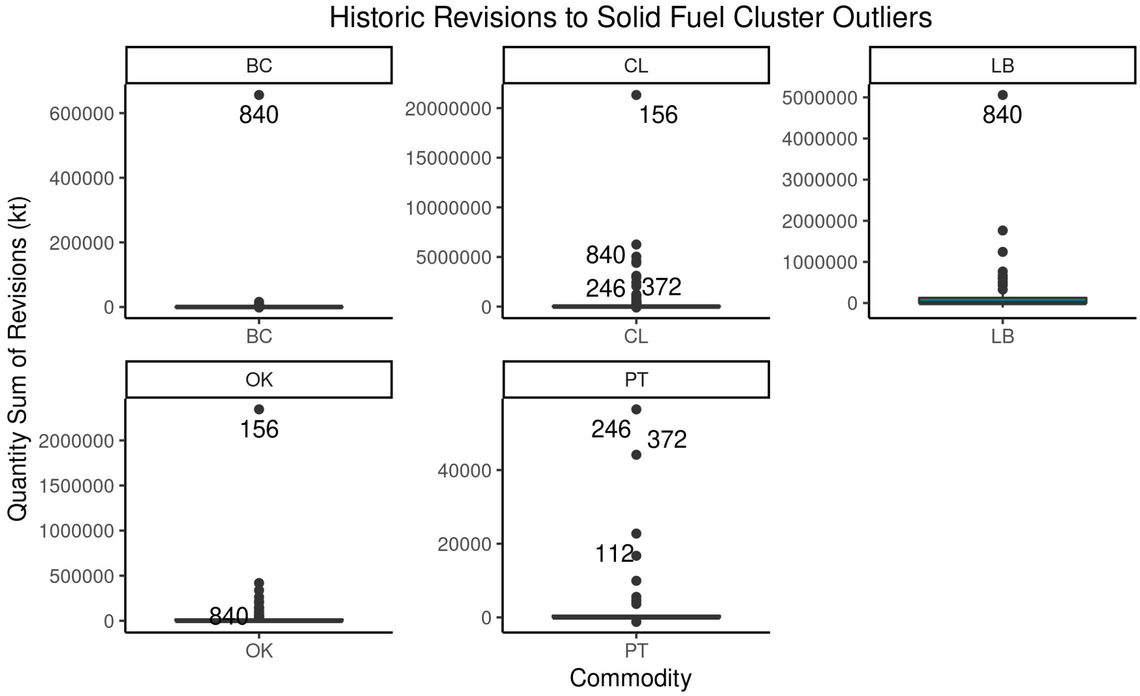

Figure 21: Outliers for Historic Revisions to Solids

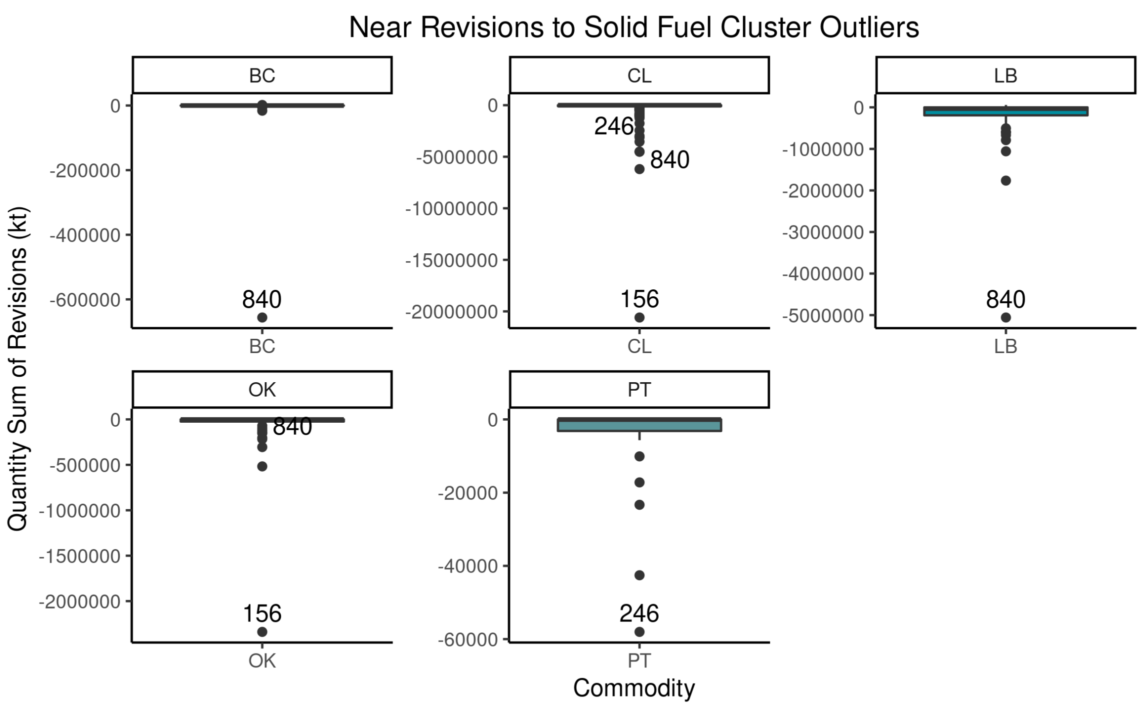

Figure 22: Outliers for Near Revisions to Solids

The boxplots for solid fuels (figures 21 & 22) reveals something interesting about these countries. The magnitudes for historic revisions are positive but for near revisions they are negative. The boxplots for near revisions look almost identical to the boxplots for historic revisions but mirrored across the x-axis. It is also interesting to note that the countries not in the main cluster for solids are the same for historic and near revisions, and so are the commodities whose boxplots they are outliers in. The revisions imply that these countries were underestimating quantities for historic entries and then have begun to overestimate near entries, since historic revisions are increasing the quantity and near revisions are decreasing it.

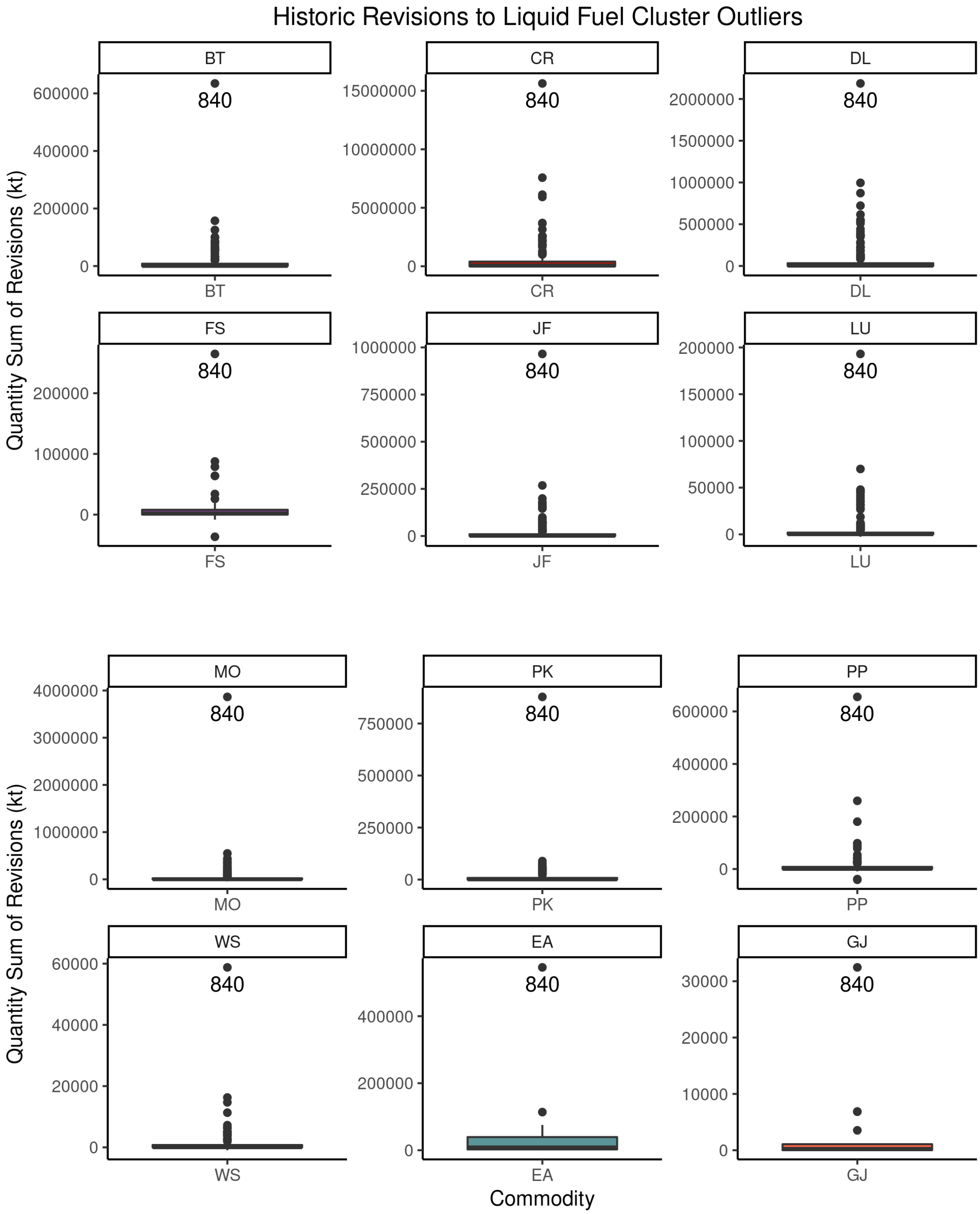

Figure 23: Outliers for Historic Revisions to Liquids

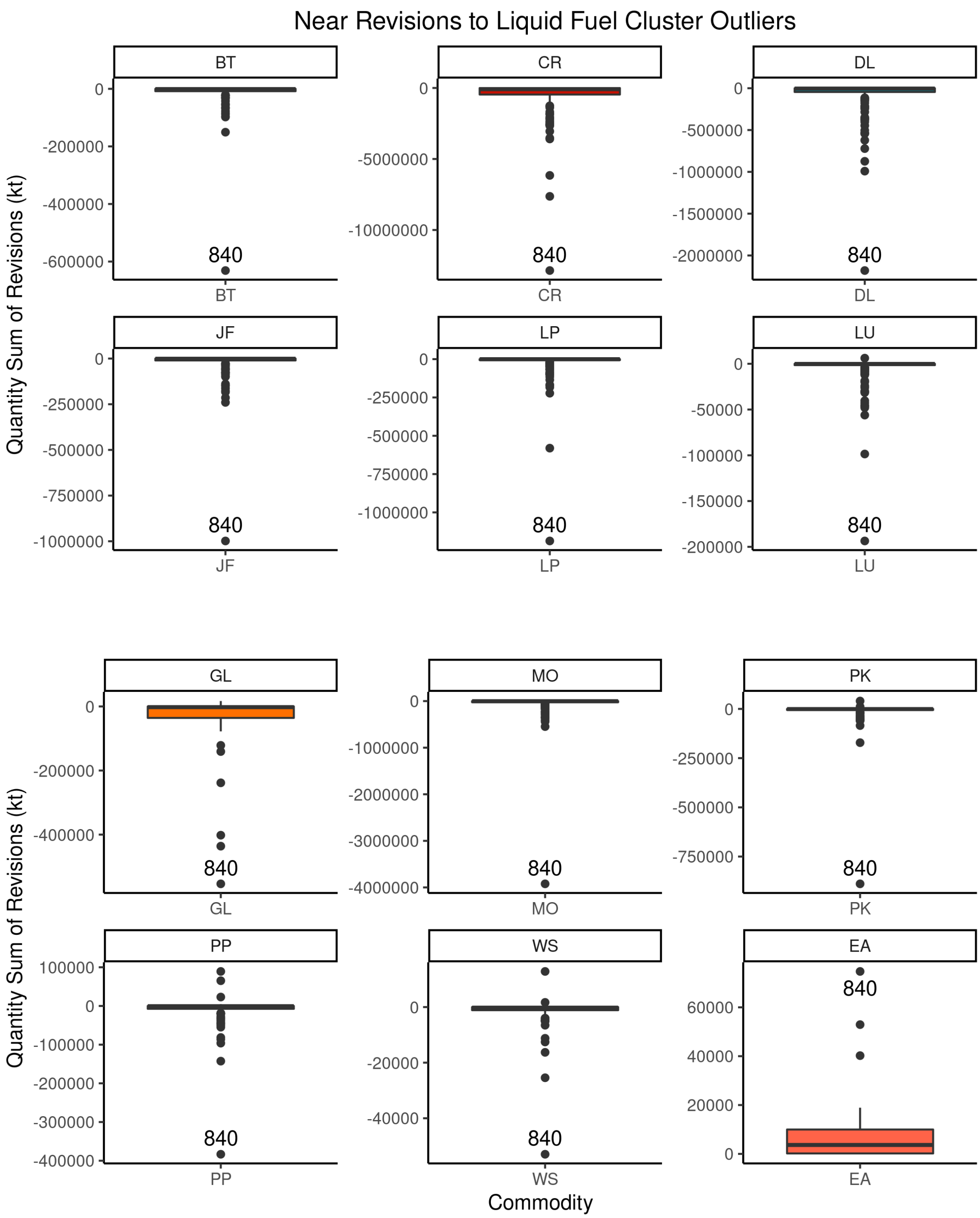

Figure 24: Outliers for Near Revisions to Liquids

For the liquid fuels (figures 23 & 24), the United States (840) was solely focused on. In the historic and near liquid fuels clusters, it was in its own cluster, extremely far from any other countries. What the boxplots for the United States indicate is that for liquid fuel revisions, it often far exceeds all other countries in magnitude. In fact, not all of the boxplots where the United States was an outlier were shown due to redundancy. Similar to solid fuels, the revision magnitudes were in the positive direction for historic revisions and the negative for near, except for commodity EA. The near revisions in these boxplots defy the trend seen in the barplot for the quantity sum of revisions by fuel group. In that plot, liquid fuel revisions for entries with an age less than ten years were for the most part positive. However, within these boxplots, the United States' revisions seem to go against this. Overall, the United States seems to consistently be making revisions to many liquid fuels on a completely different magnitude than any other country.

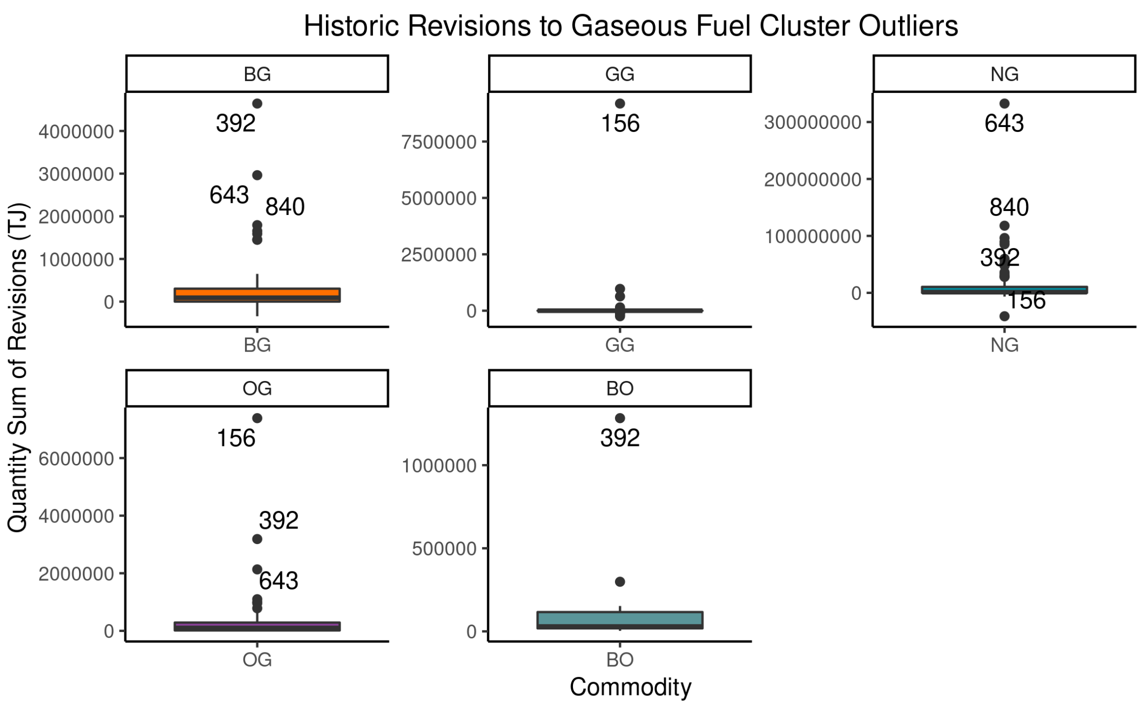

Figure 25: Outliers for Historic Revisions to Gasses

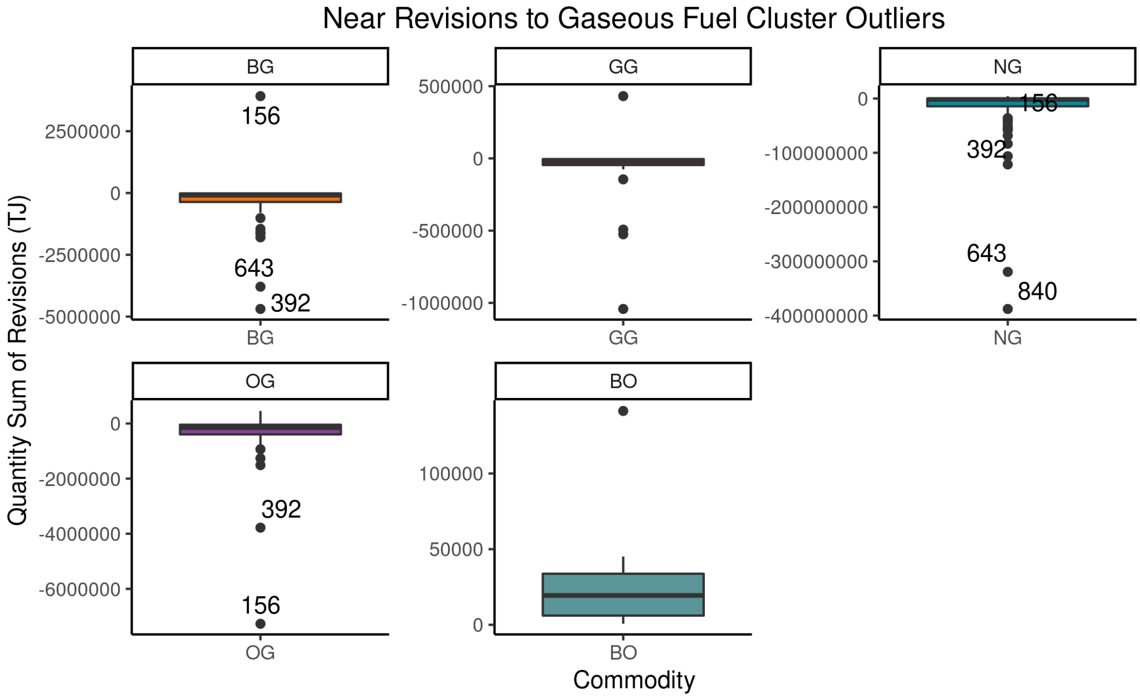

Figure 26: Outliers for Near Revisions to Gasses

Gaseous fuels show a switch from a positive magnitude to a negative when going from historic to near revisions (figures 25 & 26). It is interesting to note that in near revisions, none of the countries are outliers for GG or BO like they were in historic revisions. The near revisions show values similar to what we would expect from the quantity sum of revisions by fuel group barplot, where the magnitude was overwhelmingly negative for revisions to entries less than ten years in age.

Across all the boxplots, an interesting trend of commodities with a positive magnitude for historic revisions becoming negative when near revisions are examined can be observed. This could imply an underestimation of quantities in the past and an overestimation of quantities closer to the present, resulting in large magnitudes of revisions needed to correct this. While this analysis is limited by only focusing on countries different from the rest in revisions, it shows what the extreme end of revisions made by these countries looks like. In many cases, these countries are outliers in how they make revisions to certain commodities for both historic and near entries. This could imply larger faults in how these countries are collecting, quantifying, and reporting data.

6 CONCLUSION

The United Nations Energy Statistics Database is an important database for global energy statistics because of its temporal coverage and comprehensive inclusion of all nations. The temporal coverage can provide insights into the nature of revisions, and how they might influence estimates of $FFCO_2$. The database is revised with each release, and within just a couple years large amounts of change can occur within the data. In addition, the database shows how important a factor the age of entries are to revisions. Far more revisions happen to newer entries. This could indicate a certain level of uncertainty within the initial reporting of fossil fuel usage that is then corrected in the subsequent years to produce a more accurate value. Likewise, the total magnitude direction of revisions seems to be dependent on fuel type, with gaseous fuels being negative while solid and liquid fuels are positive in change.

Countries also play a large role in revisions. Some countries like Japan or the Netherlands will make large total magnitudes over many smaller revisions whereas other countries will make larger changes in magnitude in less frequent revisions. The way in which countries revise data for each fuel group is also different, with the United States and China being clear outliers from other countries in their revisions. These countries that produce different trends in revisions also show an interesting relationship between the age and directional magnitude of revisions. Revisions to historic data often has a positive magnitude, suggesting an underestimation of initial data that was not initially corrected for quite some time. In contrast, near revisions suggest an overestimation of initial values with the values then being quickly brought down in magnitude.

This then has implications on the uncertainty within $FFCO_2$ estimates. Uncertainty can depend on the fuel group, age, and country of the estimate. Observations in the near term are likely to come down in total magnitude while older data is likely to be brought up. It is widely accepted that the magnitude of $FFCO_2$ emissions has been growing throughout human history [1]. If this is the case, however, the difference in $FFCO_2$ emission between now and the past may not be as large as currently thought. However, within each fuel type this could be different. For gaseous fuels the values may be an overestimate while solid and liquid fuels underestimate fuel usage. The uncertainty for countries can vary wildly and some countries need to be analyzed entirely on their own. This all implies that looking at the uncertainty of $FFCO_2$ estimates as a whole may not produce as accurate a result as looking at uncertainty by fuel group or country.

As our methodology for the collection and reporting of energy statistics is improved, the need for revisions to past observations substantially grows. Each new expansion to our knowledge and understanding of energy statistics calls into question the certainty of our past observations. These statistics are, to a certain degree, approximations made based off of current framework and insight. As new information is collected and knowledge gained, previous observations must reflect this through revisions. The necessity for revising data is not an indicator of failings but a demonstration of growth and improvement in the understanding of the topic. With the rising need for accurate CO_2 emission values, the revision of past estimates provides insight into the uncertainty of the data. Through better classifying and understanding these revisions, the ability to produce accurate and reliable CO_2 estimates can be obtained and hopefully provide a pathway to combat global climate change.

References

[1] R. Andres, T. Boden, F. Breon, P. Ciais, S. Davis, D. Erickson, J. Gregg, A. Jacobson, G. Marland, J. Miller, T. Oda, J. Olivier, M. Raupach, P. Rayner, and K. Treanton, *A synthesis of carbon dioxide emissions from fossil-fuel combustion*, Biogeosciences **9** (2012), no. 5, 1845–1871, DOI 10.5194/bg-9-1845-2012.

[2] R. Andres, T. Boden, and D. Higdon, *A new evaluation of the uncertainty associated with CDIAC estimates of fossil fuel carbon dioxide emission*, Tellus B: Chemical and Physical Meteorology **66** (2014), no. 1, 23616.

[3] S. Chartier and L. Morissette, *The k-means clustering technique: General considerations and implementation in Mathematica*, Tutorials in Quantitative Methods for Psychology **9** (2013), no. 1, 15–24.

[4] D. Gilfillan and G. Marland, *CDIAC-FF: global and national CO_2 emissions from fossil fuel combustion and cement manufacture: 1751–2017*, Earth System Science Data **13** (2021), no. 4, 1667–1680.

[5] Statistics Norway, *Report from the Oslo Group on Energy Statistics*, `https://unstats.un.org/unsd/statcom/doc13/RD-OsloGroupReport.pdf`. Accessed April 20, 2021.

[6] J. Tukey, *Comparing Individual Means in the Analysis of Variance*, Biometrics **5** (1949), no. 2, 99-114.

[7] United Nations, *International Recommendations for Energy Statistics (IRES)*, Statistical papers, UN, 2018.