**Energy Mix Transition Sketch**

Focus of paper: ‘states of change’ – i.e. how energy tendencies manifest in the U.S. states in 1990 and 2016.

 Provisional questions:

To what extent is disruptive change evident in the states? (We’ll need to set a threshold for each attribute.)

Are substantial changes in energy jobs associated with traditionally coal and nuclear-intensive energy mixes? – to look at after thinking about more and starting with transitions

To what extent are Democratic states associated with more low carbon energy mixes (nuclear + RETs) and/or growth in RET jobs?

* Two models: mix as dependent and growth (and stratified for both) as dependent

In which states is the presence of an RPS most strongly associated with new, renewable energy mixes?

* RPS as explanatory

Variables of interest: changes in the energy mix, the scale of consumption, the # and type of energy jobs, presence of certain policies (RPS, etc), political orientation (based on popular vote for president in the national election), carbon emissions, etc.

What is most correlated with our variables of interest? For example, how about GDP, Energy Intensity of states?

Not only are we interested in changes in energy mix, but what about proportions of energy mix/profiles?

A lot of this is categorical. More than a model for how much the energy mix changed, we’re really curious about whether or not renewables increased or decreased. And the dep. variable could be increase/decrease in wind, solar, renewables as a whole, etc.

I definitely will have to review models from my Categorical Data Analysis textbook to think about how to model these attributes. Policy presence is obviously binary explanatory, political orientation is binary explanatory, #/type energy jobs is also or at least can be made that way, etc.

OH Thursday: 9-12noon

1/12/18 Phone Call:

Start with the data for energy transitions and try to determine how to categorize them

On one project they found the cutoff to be 15% increase share in renewable and at least 100% increase in major renewable category (representing real investment not a shift because of some natural resource or other reason)

Can also try mapping the shifts by category and overall and checking percentiles and doing different categories that way

To start analysis, we’re trying to explore how different decisions for energy transition change the composition of “disruptive” change. Some ways to try

* + 15% overall increase share in renewable and at least 100% increase in major renewable
  + Rank the increase percentages for overall shares and break into categories (2 levels, or multilevel with different classification techniques for labeling)

Would be interesting to get ranks for changes at each level as well as overall. Rank by absolute value, but then add sign. We’re trying to track most disruptive.

EIA SEDS data linked below can be downloaded and assessed

Data Sources to Review:

EIA SEDS data: <https://www.eia.gov/state/seds/>

Consumption Data

* Consumption Overview: Primary energy source and end-use sectors
  + State by state for total energy, itemized fossil fuels, nuclear, renewable energy (not itemized), end use sectors, electricity import
* Energy Consumption Estimates for Major Energy Sources
  + Coal, Natural Gas, Petroleum stratified, nuclear, hydro, wind
* Primary Energy Consumption Estimates
  + Includes more stratified renewable energy estimates
* Total end-use is stratified
* Electric Power Sector Consumption Estimates
* Production sets also of interest

**January 17, 2018:**

Downloaded use\_US.csv. Has US data for different codes. I think this is nationwide consumption for 332 codes/descriptions

Downloaded StatsCons\_Codes\_and\_Descriptions. This has been helpful to go through to think about how to use the data downloaded. I’m going through the list looking for total consumption categories, highlighting them in an excel to use for state transitions data. The idea here is that when I find the dataset that has the states consumption in different codes/descriptions I can remove rows that don’t match categories of interest and then proceed with the analysis. At first, I wasn’t sure which units to use (thousands of barrels and BTU appear for many total consumption categories) but I’ve decided to stick to BTU. It is always present for consumption categories, regardless of source. There are some areas I want to learn more about. For example, total jet fuel consumption a sum of the other jet fuel consumption categories (like kerosene) or is it something different? This can be verified, and I plan to verify things like this. Another point of interest is verifying whether fossil fuel total consumption is a sum of all the categories I think it is. If not, I’ll have to look for more information on the data than I’ve been able to find so far. I haven’t found anything that described the codes/descriptions in much detail yet. The point of understanding these better is to make sure the analysis is correct (ie not arbitrarily taking ratios for categories since this could introduce a lot of noise into a probably already noisy dataset)

Update: last 3 characters TCB or TXB indicate Total Consumption BTU or Total end-use Consumption BTU. That will be the way to determine which rows to keep

<https://www.eia.gov/state/seds/sep_use/total/csv/use_csv_doc.pdf>

See page 7

Downloaded Complete\_SEDS which has state codes and all codes/descriptions. So 1960-2015 is 55 years x 50 states x 930+ codes is around 2 million (data says 1.6 million rows so rough estimate). There are some 0 data entries. For starters, load into data frame and drop all the rows without the 3 characters named above, then see what’s there to see where 0’s are, and other things like that. Then can decide how to shape the dataframe/matrix for analysis. Keep the US consumption data separate for now. Think about that more later/ find out what’s in it later.

Made total consumption and total final end use consumption dataframes out of the overall dataset, Butane units

Total Final End Use:

ARTXB AVTXB CLTXB DFTXB ESTXB GETXB HYTXB JFTXB KSTXB LGTXB LOTXB LUTXB MGTXB NGTXB P1TXB PATXB POTXB RFTXB SOTXB TETXB TNTXB WWTXB WYTXB

Total Find Consumption

ARTCB AVTCB BMTCB CLTCB DFTCB EMTCB ESTCB FFTCB GETCB HYTCB JFTCB JKTCB JNTCB KSTCB LGTCB LOTCB LUTCB MBTCK MGTCB MMTCB NGTCB NNTCB P1TCB PATCB PCTCB PMTCB POTCB RETCB RFTCB SFTCB SOTCB TETCB WWTCB WYTCB

length(which(state\_total\_fconsb$Data\_Status != "2015F"))

length(which(state\_total\_consb$Data\_Status != "2015F"))

both = 0 so I dropped this column. All data for consumption is 2015 status

length(which(state\_total\_consb$Data <= 0)) = 10,309

length(which(state\_total\_fconsb$Data <= 0)) = 8,065

which means 10.8388% and 12.04163% of data entries for state total consumption and state total final end-use consumption is <= 0 respectively. Are these entry errors? Note that there are very few negative values and they only occur in the state\_total\_consb set (44 instances)

Use state\_total\_consb[(which(state\_total\_consb$Data < 0)),] to see the entries

95,112 rows in total\_consb and 66,976 in total\_fconsb

I want to do things like

* + Track changes in energy type in each state
    - Set a criteria that picks out which 2 year periods have dramatic changes (could be noisy). Maybe, for each year, call these “dramatic events” or something like “disruptive energy transitions” and we can see which years have the most of these and if they have characteristics (partisan lean, economic growth, etc.)
    - Do this for renewable vs fossil fuel, individual areas, etc.

First, I have to get 1960 and 2015 data and figure out a way to get the dataframes organized for the overall growth comparisons

Stc\_comp and stfc\_comp store these. What I want to do is write a function that takes in an MSN code, then outputs chart with the % changes for each state for that category

Also, function that takes in a state and outputs % changes for each category

I have to think about the data a bit more and how to get these functions working. Will work on next time.

Use these links when starting again (just in case):

* + <https://www.eia.gov/state/seds/sep_use/total/csv/use_csv_doc.pdf>
  + <https://www.eia.gov/state/seds/seds-data-complete.php?sid=US#stateSelection>

**January 19, 2018**

Downloaded tidyverse -

<https://www.tidyverse.org/learn/>

<https://www.rstudio.com/resources/cheatsheets/>

<https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Tidyverse+Cheat+Sheet.pdf>

Print these for manual reference while coding until memorized, known very well

Changed column name for Data to Billion\_Btu for referencing consumption

Turns out the single observation rows is optimized for data science so in good format already. Just need to write the function that takes in the MSN code and outputs percent change, and can do things based off that function. I want it to return a 2 col. Df for state and %change cons.

For this method I want to have a dictionary that maps the MSN codes to the full word meanings of that they are so the data frame column for % change is clearly identifiable and corresponding charts, figures, etc. are clear. I have to put more thought into that

<https://stackoverflow.com/questions/24212723/return-a-data-frame-from-function>

**January 22, 2018**

The states.perc.change function requires a dataframe passed in with the correct extraction of base and comparison years. For now, this enables me to do 1990/2015 comparisons for the sectors/MSNs of choice (ie wind is top priority).

The function will return a df with the state, a string column for years compared with the MSN code in front, and the percent change for that baseline to comp year.

temp <- msn\_frame %>% distinct(Year)

temp$Year returns the Year vector and can then access the elements to make a string for base/comp

It’s probably not optimized to have an entire column for MSN Base Comp, because that’s only needed for labeling visualizations, charts, etc. later.

One way the method can be fully scalable is if it could take a data frame where it groups by state and returns a data frame with the pct changes over all the years listed in the original dataframe (has to be in order where subsequent years are compared against each other. This would be used for getting year-to-year 5 year, 10 year, etc. type comparisons. Or baselines of interest like 1990-2015.) The problem with the way I return the dataframe now is that it would return “1990-1995-2000….2015” For example if you passed in a 5 year sequence up until 2015. That’s not a problem, now that I think of it. You’d just take the string separation method and do it for each subsequent, so you’d have rows like 1990-1995, 1995-2000, etc. with MSN code in front of each of these. That repeated MSN is redundant, but I can’t think of a way around that yet.

So, for now I’m leaving this method as is. However, it needs the following two improvements

1. Take multiple years, not just 2
2. Get the years comp string to generally, not hard code each comparison (not always going to be length 2 so have to loop through the back to back to paste together)

Also of interest is having a data structure in place that maps the MSN to full Description so the string can be fully written out and makes plotting look better

Tried to test Wind use Total Consumed. 1990 has many 0 values so we’re getting Inf returned when there’s a 2015 value and a 1990 0, and NaN when 0 for 1990 and 2015

So, when passing this DF into further functions, need to extract these rows. For visualizations, can take a note that says “x # of states have 0 Btu for 1990 and 2015 MSN” and “y # of states have 0 Btu for 1990 but some Btu for 2015 and therefore a percent change can’t be calculated”

The problem for the regression modeling is that we now have states that are excluded? So, one way around this is to classify states into the following categories:

1. No change
2. Change from 0
3. Positive % change
4. Negative % change

Or something like that. Unfortunately, we lose a way to relatively compare states in the ‘2’ category, unless they’re compared against themselves.

So, it’s possible that the model can look at ‘2’ states alone, and ‘3’ and ‘4’ states together. What to do with ‘1’ states?

Next Steps:

MSN mapping

Take in df to produce some visualizations

Think of further methods to write

Generalize current methods

**January 23, 2018**

Finalized the method for getting the %change and rel. change df

Worked on the visualizations methods

The code is commented well with ways to proceed for the plot function and different ways to proceed. I will be able to discuss those more in depth in the work log next time I work on this

**January 24, 2018**

Working on the MSN mapping right now

* + Easy solution: Just read in, do TC|TX grepl and B for units and only 57 codes. Can now use these as a basis for further exploration
  + Also, pretty easy one line command for filtering and then accessing Description field of return to get a string for printing to plots

Next time working on plotting methods and also going to do some exploration of which fields add up to others, etc. by subsetting based on codes, seeing which are related, etc. I can do this on a state level and on the US level to see if it’s consistent

**February 4, 2018**

Sent following email:

Update/questions:

I'm testing my method on the Total Wind Consumption and noticed there are only 4 states who have a % change from 1990 to 2015. There are many states that have a relative change (growth from 0 at 1990 to some value at 2015) and some states that were 0 at 1990 and 0 at 2015. How would you like me to handle this?

The way I'm currently handling it is printing which states were 0 in both the base and comparison years. Since the change from 1990 to 2015 can be huge, the plots are only decipherable on a transformation of the growth (difference of consumption at 2015 and 1990) ie a log scale. I was thinking of doing the following:

 1) Plot all the states that have a change in consumption from the base to the comparison year on the same plot

 2) For the states that have a % change (not 0 in 1990), annotate the %change above their bar on the plot

 3) Group the states by some political leaning score and group the color sequence into red/blue to add more dimensions of data to the plot

To elaborate on the 3rd point, let's say California is the most blue state. It would be in the left most position in the plot, and the darkest blue. Then all the way to Texas (maybe that's not true anymore for most red) the gradient shifts from dark blue to light blue to light red to dark red. This adds a visual grouping dimension to the plots which I think is important since bar charts can be very hard to make comparisons in (especially for so many states).

Lastly, how would you like me to handle DC and US? The EIA consumption data includes these codes. I can remove them, or just treat them like other states. Currently, I've been removing DC and pushing US to the furthest right of the plots so it can be compared against all the other states.

Looking forward to your thoughts.

Response: Keep US, delete DC, everything else sounds fine. Keep up to date on if analysis on particular segment is taking up all of time dedicated.

Some ideas for today’s work:

* + <https://stackoverflow.com/questions/28755576/add-text-to-ggplot>
    - Can add vectors to annotate and factors have numeric values. Each state has its own index so that can be passed in for vectors, with a corresponding growth + .015 or some small perturbation for spacing that allows a nice clear show of the annotation. It’s really easy to see other posts on annotate that give an idea of how to add the right space at a y-level. So, x and y will be passed in this way (processed through the subset with perc. Growth). There will need to be some trickery done with indexing to get it all right, but it’s definitely do-able
  + Get the annotate part working. Get the rel. changers and perc. Changers into one plot, print the states not on the list. <https://stackoverflow.com/questions/11938293/how-to-label-a-barplot-bar-with-positive-and-negative-bars-with-ggplot2> is a good reference for having pos. and neg. changers possible
  + Once the method is working in the sense that it produces visualizations where comparisons can be made, and inference can be drawn, it’s easy to run through MSN codes to get the results
  + Some of the features like organizing states from blue to red and coloring/gradient, etc. need to be done later on in the process. No data with partisan score yet, for example.

Once this is in bare bones working order, we’ll want to get some results for some of the important codes and then start thinking about models. So, the goal for this week is to finalize the plot, have figures saved for analysis, and get thinking about the models. Will need to talk to Prof. about data of interest like partisan score, the policy she mentioned, etc. How do we integrate jobs? Etc.

Update:

Completed the method that takes in frame, maps the comparison through MSN code, has base years, etc. The data structure output can be used for all plotting purposes.

Working on plots method:

The rel\_growth df have ‘inf’ percent change whereas movers has finite perc. Change. Both have some relative change (from a 0 base or some other base)

First thing the method does is print out a list of states w/ 0 cons. In both base and comparison year. Fine for now

The plot will be all the rel\_growth and movers with US last (furthest right for now, but that could change). The idea is for now, no order other than alphabetic (but with political parity there will be an ordering and color gradient to represent another dimension of data at some later point). The scale will be log change, and the states that have %change will have an annotation for it, whereas others will have no note at all. Some note on the plot can state that states without annotated percentage had 0 consumption in base year 1990.

Data wise, the rel\_growth and movers should be in the same data frame. And, the indices should be stored for the ‘movers’ so they can be referenced for annotation. Need their percent change to be stored with indices in two vectors that will be used for the annotation call.

Do we want to normalize these values so that they are a % of total consumption? (We have the total consumption and total final end use, so if they fall into either they will be normalized by respective one.) This could help with the meaning of the charts. Because the log scale imparts an information loss. The relative growth is really what matters, but the scale is hard to interpret.

Resources for later work:

<https://stackoverflow.com/questions/11938293/how-to-label-a-barplot-bar-with-positive-and-negative-bars-with-ggplot2>

<https://stackoverflow.com/questions/28755576/add-text-to-ggplot>

<http://ggplot2.tidyverse.org/reference/annotate.html>

I’m leaving off with that thought, for now. Want to work on the annotation next time, the normalization to see if the plot looks better (so back to original method), and getting things together for running through analyses.

Another thought is to take time off the bar plots, and focus on tables since there are so many comparisons to be made. Just printing the data frames through to tables with mappings for Inf and non finite, and putting it together so comparisons can be made.

The plots are going to be too tough to get meaning across for because so many levels. The map is better for that comparison anyway, and the color gradient on the map will represent parity, and annotations can directly be on top for growth, or density dots, etc. Lots of options with maps. So, don’t focus too much on plots, get together tables, get the things output so we get an idea of what’s going on in dif. States with different sources, and look at normalizing for some added context. This can help us get to modeling faster, and the visualization work should be focused on maps, not plots. Especially not bar plots.

**February 6, 2018**

It’s tricky. Not sure how to represent the changes the right way. The normalization by change in total consumption for a category might not be the best way since it’s just the %change the particular category makes of overall change.

change$Change/normalize$Change

I have to think about this more.

**February 9, 2018**

I can output the data in tables which is good for just getting some results off the bat. We can look at increase/decrease, %increase/%decrease, and normalized values (percentage of total consumption or total end use consumption for that particular state/entity). This gives us a good idea of which states are changing the most, and which are changing the most relative to their own composition (ie California will probably have large changes in each renewable category, but they’ll also have large changes in consumption so as a % of consumption it might not be as significant a change as let’s say Nebraska).

Getting the printing of the table is going to be tricky, so make sure this is the right thing to do before spending a lot of time on it.

Another thought – maybe instead of normalizing by total consumption we normalize by population for 1990 and 2016? Or different normalizations to add more context? Tricky.

**February 12, 2018**

Maybe a composition chart for each state and the US? This shows in 1990 what composition of energy sources were (out of total energy) and then in 2016. So it shows how states change related to the US, and by total final end use and final consumption. And it also shows it over a dimension of time. That is definitely something worth looking into. The amounts of change in terms of units of energy are really hard to display the right way visually. That’s more of an analytical concern, I think, which is best seen through a table.

And, I think that what’s important in the composition charts is whether the total consumption and total final end use consumption increased. If not, just a note that says it decreased. Highly unlikely any decreased. There are certainly % changes of total consumption that are worth looking at. For example, how much was the %change of a certain source consumption divided by %change of total consumption? It’s getting hard to figure what the right units of comparison are, but this seems promising right now.

**February 19, 2018**

Notes on phone call:

Use factor level increase instead of % (useful in Prof. Araujo’s opinion)

Comparison across technology, not just states. Index everything to base year. (Assign base year 1, and then do a 2016 relative to base year)

Non-action

Slow changing

Changers

One version – per capita

One version – absolute, but indexed based on factor level increase

SEDS should have population for 1990. If no data on this from set, use Census Bureau.

How do you separate laggards from others? How do you separate fast changers from changers?

Hypothetical layout:

Section 1 – what are laggards, non-movers, and prime movers? For a major category like wind, maybe do top 5 movers. Then see what was happening in factor increase, % of total consumption in base and final

Section 2 – Looking at these top movers, what characteristics do they have in common? Maybe job growth, political orientation, other things are lining up in some way?

End Use and Final End Use – where do substantial differences exist? Principally in transport and losses.

Transitioned coding to python/pandas. Took less than 30 minutes to get the data frames the right way. So much easier for me.

Function that returns a data frame for a given code

* Column that labels as non-mover, slow-mover, and prime-mover (0-0, 0-some number, number – number)
* Prime movers only
  + Column for factor level increase (2015 amount / 1990 amount)
  + Column for consumption per capita in base, consumption per capita in next, change in consumption per capita
  + Column for consumption as proportion of total consumption category in base, next, and change
* Slow-Movers
  + Column for consumption per capita in new
  + Column for consumption as proportion of total consumption

Sort according to consumption as proportion of total consumption in the new. These are probably the ones we’re most interested in. Then we can look at tables for top 5, mid 5, bottom 5, etc. Or, when we know how to cluster these we can take 5 representatives and see what they have in common at later stages as more data is added.

Work on the data frame calculation method first, then worry about other methods like plotting (maybe won’t even work on that).

The data frame calculation method is for printing out the master tables. One table for each category. Obviously, the US is a prime mover each time. Will move to bottom of table, I think.

Good progress.

Need to figure out formatting of table better. The method is working pretty well. Plotting just needs work.