Deriving Estimates for the Energy Consumption of U.S. Residential Space Conditioning Using Seasonal Datasets

W. Gregory Lawson[1]

[1]U.S. Energy Information Administration, 1000 Independence Ave., SW, Washington, DC 20585

Abstract

We wish to estimate the end-use energy consumption for U.S. residential space conditioning (i.e., air conditioning and space heating), which is markedly seasonal. Essentially all extant estimates can be traced back to the Residential Energy Consumption Survey (RECS), a nationwide survey conducted periodically by the U.S. Energy Information Administration. This is a tribute to the value of RECS but also an indication of how difficult the estimates are to validate externally. Here we pursue an idea for an "external-to-RECS" estimate of space conditioning consumption: we seek to synthesize monthly residential energy consumption data, which is the total consumption from all residential end uses, and monthly weather data, which should only relate to the space conditioning components. Though the idea is intuitive, the results are sensitive to decisions made throughout the process: for the weather data, we work with population-weighted degree days data, which are sensitive to the base temperatures used in their calculation; monthly consumption data is the product of reporting from utilities, which means the data likely contain unknown time lags. We use simple linear regression within a framework of brute-force optimization to find the "optimal" base temperatures and time lags to use, and hence yield our "optimal" estimates for the consumption due to space conditioning. In the present study, we restrict our analysis to cover only natural gas consumption and space heating estimation. We find the method produces estimates that are quite reasonable; we even find a surprising result implying that some historical RECS results may have been obtained in a potentially inconsistent manner.

Key Words: energy, energy consumption, seasonality, Residential Energy Consumption Survey, RECS

1. Introduction to Problem

An important problem within the energy sector is estimating the energy consumption of various end uses from measurements of total energy consumption. Rarely are direct measurements available of specific end uses (e.g., lighting, air conditioning, refrigerators), so one must inevitably employ statistical modeling to defensibly solve the "unmixing problem" of going from a measured total to estimates of the summed components comprising that total. The Residential Energy Consumption Survey (RECS), conducted periodically by the U.S. Energy Information Administration (EIA), is the main vehicle for producing national end-use estimates within the residential sector (EIA, 2017). There is a dearth of independent end-use estimates, a tribute to the value of RECS but also an indication of how difficult its estimates are to validate externally. Broadly, it is our great interest to find any "external-to-RECS" estimates for end uses that can be constrained by publicly available data. Within this study, we focus on space conditioning, both because heating and cooling are two very important end uses, and there seems an intuitive line of pursuit given their marked seasonality. The particular candidate estimate explored here is that of space heating fueled by natural gas, since the number of end uses for natural gas is much smaller than for electricity. Hence, we ask the question: can we estimate consumption for natural gas space heating from monthly datasets of natural gas consumption and weather, *without* referring to RECS data?

This is an intuitive idea, for datasets that capture seasonality in residential fuel use will inevitably show seasonal peaks in usage: wintertime space heating leads to wintertime peaks in fuel consumption, and summertime cooling leads to summertime peaks in electricity usage for air conditioning. If we could compare the consumption within those peaks to the total consumption over a year, then we would have an estimate for the percentage of total consumption used for either space heating or air conditioning. Do we have information sufficient in coverage and precision to make a meaningful estimate in this manner? Despite the sensitivities to unknown parameters like degree days temperature bases and utility-reported time lags, we believe we are able to contribute a useful estimate.

This intuitive idea was apparently first explored in the 1980s. In fact, it was a substantive part of a Ph.D. thesis written in 1982 (Goldberg, 1982), and it is the fundamental basis of the so-called Princeton Score-keeping Method (PRISM) for tracking and verifying energy efficiency measures (Fels, 1986). Most of this previous work focused on the individual housing unit as the unit of analysis, though Fels and Goldberg also authored a study looking at aggregate measures of natural gas consumption in New Jersey (Fels & Goldberg, 1986). Inasmuch as anything presented below is novel, it is that we are attempting to focus on national or regional estimates. This scope is more germane to externally validating RECS estimates, and it also allows us to work with publicly available, aggregated data, as opposed to the particular billing data for given housing units.

In this report, we motivate our brute-force optimization approach and show some of its results. We will employ simple linear regression over each calendar year and region of interest. We choose to focus on separate calendar years to yield estimates directly comparable to RECS results, which also cover specific calendar years. The regression results are sensitive to assumptions about the base temperatures used in calculating degree days as well as the effective time lags in utility-reported consumption data. Hence, we use goodness-of-fit measures from our simple regressions to choose the "optimal" parameters to use for each pairing of calendar year and region of interest, and we evaluate those goodness-of-fit measures for every pairing of parameter values from some *a priori* lists of acceptable parameter values. We find that our resulting estimates of U.S. natural gas consumption for space heating are in-line with some previous RECS estimates, though they also call into question some other RECS values, an unintended by-product of this research.

2. Data Used

2.1 Monthly Consumption Data

While there are several publicly available datasets that could be useful for gleaning monthly natural gas or electricity consumption, we decided to use monthly sales data. EIA collects monthly sales of electricity by sector in their survey EIA-861M (formerly EIA-826), and monthly sales of natural gas by sector in survey EIA-857. The results of both of these surveys are disseminated as part of EIA's Monthly Energy Review (EIA, 2017). The primary advantage of using sales data is that they have already been divided by sector, thus isolating our area of interest, the residential sector. While sales do not necessarily equate to actual consumption, we feel it is a good approximation of consumption since we know that there is negligible storage of natural gas or electricity in the residential sector, and there are negligible losses of these commodities behind-the-meter (i.e., after they have entered a particular housing unit).

One disadvantage to using the sales data for natural gas and electricity is that the survey respondents reporting these data to EIA are utilities. Utilities have customers to whom they provide natural gas and electricity, and not all customers have billing cycles that naturally coincide with the beginning and end dates of calendar months. Hence, when utility X reports the total residential consumption C in month M, it is often the sum of many customers' consumption totals for a period of approximately a month with a billing end date that falls within month M. This means that the value C is a sum of integral values that have potentially different time bounds, the net effect of which means that C does not necessarily relate to the weather that occurred over M but instead some combination of the weather in months M and M-1. Said another way, the "June total" could in fact be more representative of the consumption from, say, May 20 through June 19, depending on the billing cycles aggregated by X to find the June total. Further complicating matters, some utilities make an attempt to "calendarize" their data before submitting it EIA. For electricity, it is not known well how many attempt to do this nor how well they do it (C. Reynolds, personal communication, 2017); for natural gas, the process of calendarization has occurred in-house at EIA since 2010, though the need for it was identified in late 2006 (J. Wade, personal communication, 2017). We will model these effects as a single, unknown time lag within the monthly sales data, even though we know there are actually many different time lags at work here.

2.2 Degree Days Data

For weather data, we choose to follow the tradition within energy consumption analysis of working with so-called degree days. Degree days are essentially engineering "rules of thumb" for indicating when and how much space conditioning is expected to be used based on *outdoor* temperatures. Because space heating and air conditioning are actually determined by a given housing unit's thermostat behavior, building properties (e.g., insulation), and its *indoor* air temperature, we can see that degree days are a model or heuristic for linking outdoor air temperatures to an indoor process. Having such a model is desirable because there exist easily obtained, long records of outdoor air temperatures, so it makes the ensuing modeling based on degree days much easier to complete. Further, as defined, degree days can conveniently be summed over time periods like months or years.

Heating degree days (HDDs) are the sum of degrees (°F) by which the daily average temperature at a given location dips below a set base temperature, whereas cooling degree days (CDDs) are the sum of degrees by which the daily average temperature at a given location exceeds a set base temperature. Canonically, the base temperature is assumed to be 65 °F for both HDDs and CDDs, but since degree days themselves are simply a model for indicating indoor space conditioning use, we immediately recognize that the best models may have a different "optimal" parameters. We will model HDDs and CDDs as if there were a single, unknown base temperature for each, not necessarily equal, that is representative over the region of interest, even though we know that each housing unit included in a given region could very well have its own thermostat behavior and building properties that lead to many different base temperatures being present simultaneously in a region of interest. Note that HDDs and CDDs in practice often have different base temperatures; many housing units have a range of weather conditions, a so-called “dead zone,” over which their residents choose to neither heat nor cool.

We will use daily population-weighted degree days data as made available by NOAA’s Climate Prediction Center (CPC, 2017). These data were calculated using a base temperature of 65 °F for both heating and cooling. We use population-weighted data because we are trying to find a national estimate comparable to RECS Whereas traditional degree day measures are valid only at a given point location, population-weighting many such measures within an area of interest gives a single value that characterizes that area. We use daily data for two reasons: first because we will need to insert known time lags into the weather data to account for the unknown time lags in the monthly sales data, and second because we will need daily data in order to attempt to change the base temperatures of population-weighted data that were originally evaluated with respect to a base temperature of 65 °F.

3. Methods

3.1 Simple Linear Regression

To estimate the portion of total natural gas consumption attributable to space heating, which is modeled via HDDs as being directly proportional to outdoor temperatures, we turn to simple linear regression. For any given calendar year and region of interest (e.g., the continental U.S. or a given U.S. state), we will have twelve monthly values of consumption from the EIA data, and we will calculate twelve monthly sums of population-weighted HDDs from the daily CPC data. A first step to estimating the fraction of consumption attributable to space heating is associating space heating with time-varying component (i.e., the slope times the montly HDDs) in our linear regression, and associating everything else with the remaining time-invariant component from that same regression:

$$Cons_i = \beta_0 + \beta_1 \cdot HDD_i + \varepsilon_i$$

where *i* indexes the months in this particular calendar year. Referring to this equation, the total consumption is simply $\sum_{i=1}^{12} Cons_i$, the “base load” or time-invariant consumption is $12 \cdot \beta_0$, and the consumption due to space heating is $\sum_{i=1}^{12} \beta_1 \cdot HDD_i$.

This is a very simple, intuitive approach. However, as with any regression, the results are wholly dependent on the specific data values used, and, as noted above, the data we will use are sensitive to certain degrees-of-freedom, which are not obvious how to constrain at the outset. Hence, before we can begin performing our simple linear regressions, we must first specify processes to control for these degrees-of-freedom, namely the base temperatures used in the population-weighted degree days and the unspecified time lags known to be present in the monthly utility data. Once we can control for these *a priori* unknown values, then we can attempt to solve for their “optimal” values.

3.2 Changing Base Temperature for Population-Weighted Degree Days

Our decision to use pre-computed, population-weighted degree days from NOAA’s CPC has the consequence that it becomes non-trivial to change the base temperature with respect to which the degree days were calculated. For standard degree days measures, relevant for a point (e.g., an airport for which we have a long temperature record), one can easily transform the base temperature in a given direction. For example, if we knew that there were 2 HDDs with respect to base temperature 65 °F at a given point on a given day, then we immediately know the daily average temperature was 63 °F at that point on that day. Knowing the underlying daily average temperature then allows us to calculate the HDDs with respect to any other base of interest. However, if we knew there were zero HDDs with respect to 65 °F, we do not have enough information to specify the daily average temperature, except to say that it was $\geq$ 65 °F: some information is lost in the thresholding in the definitions of degree days.

In the case of population-weighted degree days, there is no simple way to use the weighted degree days values to infer the underlying effective daily average temperature, a necessary first step in calculating weighted degree days with respect to a different base temperature. One could treat the population-weighted values as if they were point values, but this would undermine the reason for using population-weighted values in the first place. Another problem with treating population-weighted values as point measures is that one can obtain logically inconsistent results. For example, it is common to have non-zero weighted HDDs and CDDs, both calculated with respect to the same base temperature, on the same day for a large enough region, whereas this could never occur for a point: if there were 2 HDDs and 2 CDDs reported for the same day, then according to the HDDs, the daily average temperature was 63 °F, but according to the CDDs, the daily average temperature was 67 °F. Which is correct? Of course, if one had all of the original, historical temperature records and exact weights by region, then one could simply calculate the weighted degree days with respect to a different base temperature properly from first principles. Unfortunately, we do not have all of the required information at hand, and so we have to make some further assumptions in order to be able to plausibly change base temperatures.

Given a time record of daily population-weighted HDDs and CDDs (HDD_w and CDD_w) for a given region, we seek a method to infer an effective underlying daily average temperature for that region so that we can calculate the weighted degree days with respect to different base temperatures. We find that to do this, we need both an effective underlying daily average temperature as well as some effective measure of temperature variability about that average value across the region of interest. We have devised a way to solve for a mean and a variance of an underlying distribution of temperatures, which provides exactly the values HDD_w and CDD_w when the degree days definitions are integrated using the weighting by the underlying distribution. Within this process, we are essentially exchanging two pieces of information about the temperatures over an area, HDD_w and CDD_w, for two other pieces of information, the mean and variance of a distribution. Strictly speaking, this exchange only works exactly on days when both HDD_w and CDD_w are non-zero because of the information loss due to thresholding when either quantity is zero. Fortunately, this is not a worrying restriction since the larger the region of interest, the more often days occur when both HDD_w and CDD_w are non-zero (think of the continental U.S., where it is common to have simultaneously large regions in the South significantly warmer than 65 °F and large regions in the North significantly colder than 65 °F). Given that we have only two parameters to describe such a distribution, it seems parsimonious to assume that distribution is a normal distribution, since supplying any further information would require justification not provided by the data themselves.

Continuous representations of HDD_w and CDD_w with respect to base temperature B over normally distributed temperatures are $HDD_w = \int_{-\infty}^{B} (B - T) \cdot N(T \mid \theta, \sigma^2)\, dT$ and $CDD_w = \int_{B}^{+\infty} (T - B) \cdot N(T \mid \theta, \sigma^2)\, dT$, respectively, where $N(T \mid \theta, \sigma^2)$ is a normal distribution over T with mean θ and variance σ^2. Note that HDD_w and CDD_w need not be evaluated with respect to the same base temperature, but since we receive both datasets from NOAA calculated with respect to 65 °F, we are representing only one base temperature in the equations here. These integrals can be evaluated explicitly, and doing so yields two expressions that allow us to solve for θ and σ^2 for given (non-zero) values of HDD_w and CDD_w:

$$HDD_w - CDD_w = B - \theta$$

$$HDD_w + CDD_w = (B - \theta) \cdot \operatorname{erf}\left(\frac{B - \theta}{\sigma\sqrt{2}}\right) + \frac{\sigma\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{(B - \theta)^2}{2\sigma^2}\right)$$

According to the first equation, knowing the values of HDD_w, CDD_w, and B, immediately determines the mean of the underlying normal distribution. Then, according to the second equation, the total number of weighted degree days along with B and θ can determine the variance of the underlying normal distribution, though this second equation must be solved numerically. If either HDD_w or CDD_w is zero, then θ is still determined, but σ^2 is undetermined since determining the width of a distribution requires at least two definite pieces of information; an inequality will not do. We use reasonable assumptions informed by days with both non-zero HDD_w and CDD_w to fill in estimates for σ^2 on days where one of them is zero. In this way, we are able to use daily records of θ and σ^2 to change base temperatures for HDD_w and CDD_w.

As a demonstration of how sensitive HDD_w monthly totals are to the base temperature used, Figure 1 shows how the monthly totals vary for the state of Maryland in April 2015. The left panel shows a time series of daily average temperatures (the θ parameter of the underlying distribution of temperatures) for Maryland based on its record of HDD_w and CDD_w. We have singled out two particular base temperatures, 58 and 65 , both completely reasonable values to use as a base temperature. In the right panel, we show the resulting monthly sum of HDDs as we vary the base temperature, both treating the temperature record as if it were for a point (dark blue) and treating it as if it were for an area (light blue). As we have highlighted, the monthly total of HDDs for B = 65 is 312 , whereas for B = 58 , the total drops to less than half of that. Note that the curves for treating Maryland as a point and as an area are not too different. This is because Maryland is a relatively small state without large elevation differences or other features that would affect temperature, so the underlying distribution of temperatures across Maryland on a given day does not often have a wide variance.

3.3 Accounting for Unknown Time Lags

As we have discussed above, the monthly datasets of natural gas and electricity consumption are bound to contain the effects of unknown time lags, the net effect of which is to call into question the actual time coverage of any given month's data. Does June's natural gas consumption value really cover June 1 through June 30? If we are going to perform regression between consumption values and weather values, it is ideal for the datasets to correspond to the same time periods. This means we will have to attempt to model and estimate the unknown time lags. In reality, since the consumption

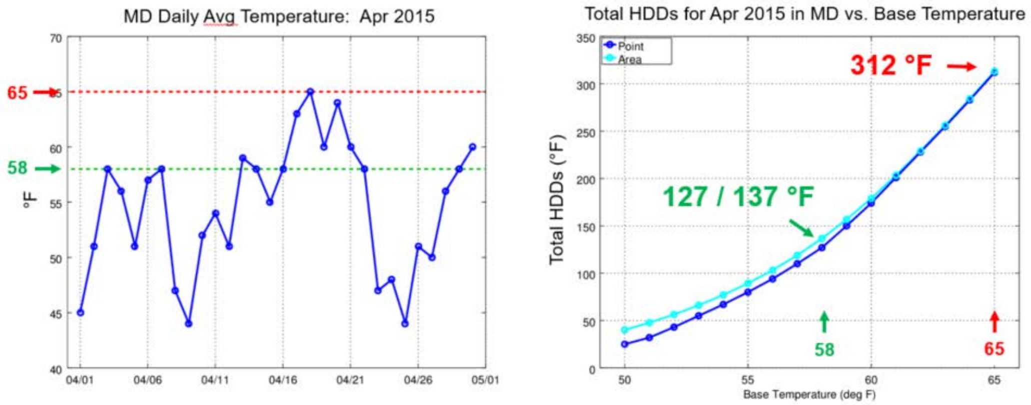

Figure 1: The sensitivity of monthly HDD totals to the assumed base temperature for Maryland in April 2015. The left panel shows the daily average temperatures ( ), and the right panel shows the monthly HDD totals as base temperature varies.

values are aggregated values from utilities, and since utilities tend to have customers staggered over many different billing cycles, there is no single time lag value we are looking for; however, for simplicity at this stage of this research project, we will model the spectrum of lags as though it can be modeled by a single, “best” lag.

Unfortunately, we do not have an easy method to alter the monthly consumption data under the effects of changing whatever lags are present. Specifically, there is no available *daily* representation of the consumption data, and we can think of no defensible methods of realistically attaining a daily version of the data from the monthly values. Hence, to control for time lags, we instead turn to inserting known time lags into our daily weather data before summing them into monthly values, with the expectation that the resulting regression goodness-of-fit statistics will help reveal the “optimal” time lags to use.

As a demonstration of how sensitive monthly degree day totals are to the inserted time lags, Figure 2 shows different monthly aggregations of daily CDD_w values for various time lags ranging from 0 days to 20 days. As is clear, it is easy to alter the shape of the seasonal peak, and as such, it is easy to change results of our regression problem.

3.4 Brute-Force Optimization

Now that we have specified processes to control for the degrees-of-freedom that we have identified in our datasets, we are ready to return to our simple linear regression. We have chosen to let the goodness-of-fit statistics from the regressions decide the optimality of our unknown parameters. Hence, our updated regression model approach now follows this equation:

$$Cons_i = \beta_0 + \beta_1 \cdot HDD_{w,i}(B, \tau) + \varepsilon_i$$

where B is the base temperature of the population-weighted degree days, and τ is the time lag inserted into the daily degree days data. There are many goodness-of-fit criteria one could select for such a straightforward exercise, and many of them are essentially equivalent. For this work, we focused on the root-mean-squared error (RMSE) between the regression model’s predicted monthly consumption and the actual reported monthly consumption. As such, our optimization problem is one of minimizing RMSE.

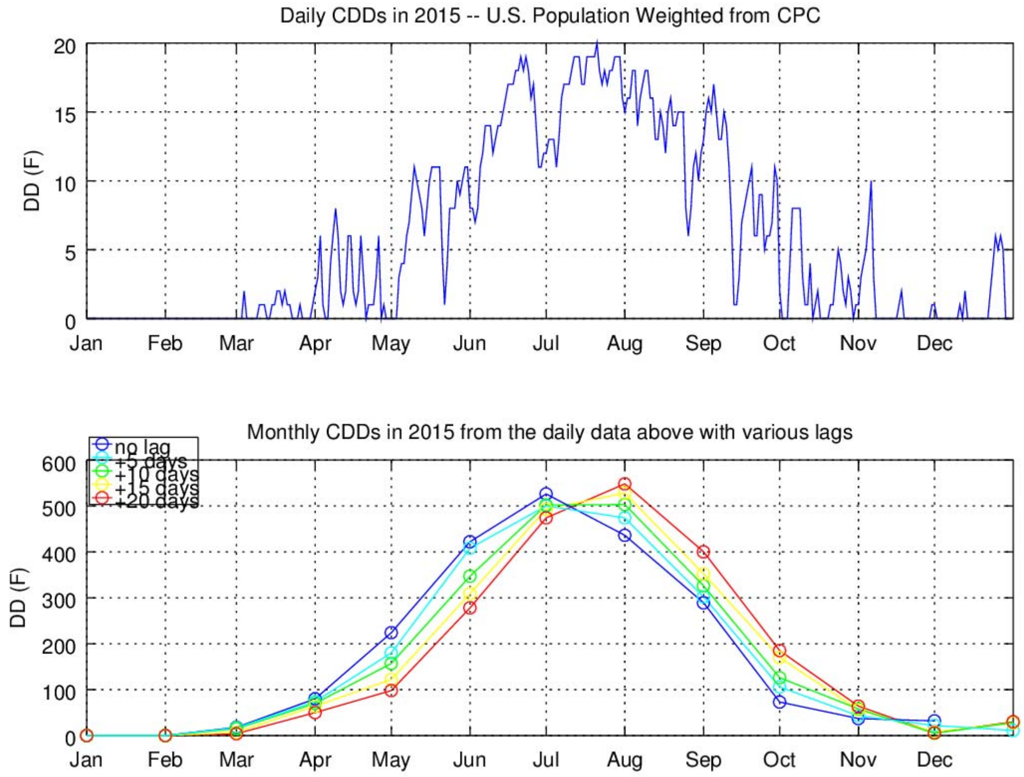

Figure 2: The sensitivity of aggregated monthly CDD$_w$ totals to the insered time lag. The top panel shows the daily CDD$_w$ data for the continental U.S. (CONUS) in 2015, and the bottom panel shows the resulting monthly totals from various assumed time lags: the blue circles show the monthly sums for no time lag (0 days), the light blue circles are from a lag of 5 days, the green circles are from a lag of 10 days, the yellow circles are from a lag of 15 days, and the red circles are from a lag of 20 days.

Based on prior expectations, we set the initial search range for the parameters as follows. For base temperatures, we assume that the optimal value for HDD$_w$ is somewhere within [50 °F – 65 °F]. For time lags, we assume the optimal value for τ is within [0 days – 15 days].

4. Selected Results

4.1 Minimization Surfaces

We ran this optimization procedure over all calendar years from 1991 to 2015, and we applied it to all 48 U.S. states in the continental U.S. (CONUS), CONUS itself (not including D.C.), as well as the 9 Census Regions. These choices were dictated by the data availability, particularly of the daily population-weighted degree days. Even with a simple, intuitive problem such as this one, there are many different results one could showcase in a report such as this. We choose to focus on a few key outputs to give a flavor of how the method is behaving. The first such demonstration is a glimpse at how the brute-force optimization procedure behaves.

Figure 3 shows the "minimization surface" when using CONUS data from 2015. The figure shows two equivalent depictions of the RMSE values spanning the ranges of tested values for base temperatures and tested values for time lags: the left side shows shaded cells, and the right side shows smooth, continuous, labeled contours. Either depiction shows that there is in fact a “global” minimum pair of base temperature and time lag. It is notable that the minimum occurs within our prior expected ranges for these parameter values. It is also notable that the minimum occurs in a somewhat smooth basin that is also somewhat broad, the implication being that the exact numerical values for the optimal parameters probably have sizable error bars. Said another way: though in this case shown in Figure 3, the approach found the optimal values to be 58 °F and 2 days, the noise involved and the inexact nature of regression should probably lead us to treat all values within the range of 57 or 59 °F and 1 to 3 days as “optimal.”

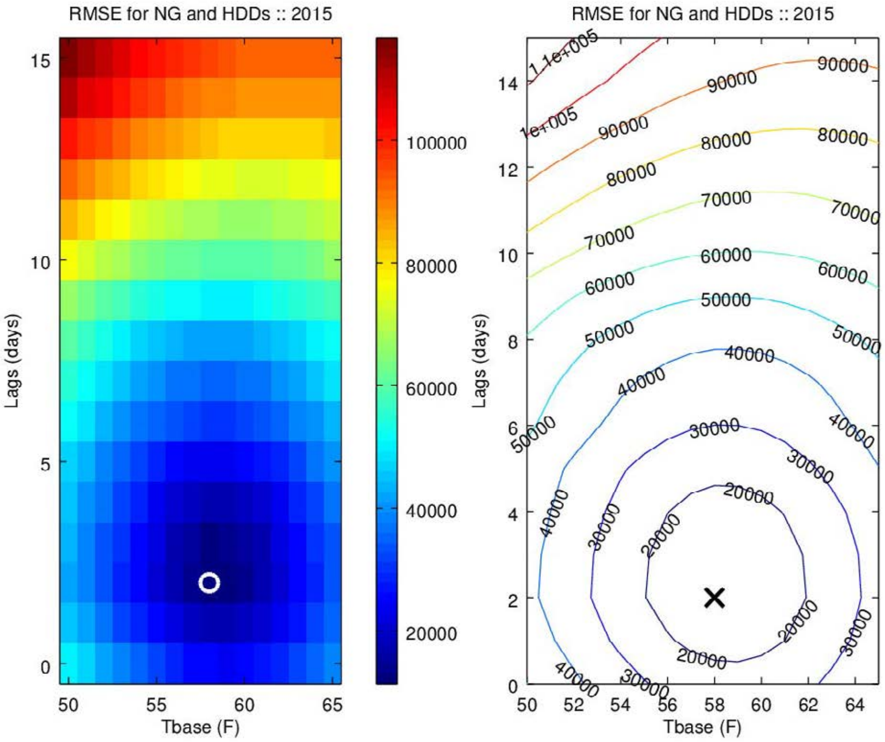

Figure 3: Two depictions of the same minimization surface. The x-axis spans the tested values of base temperature for HDDw (°F), and the y-axis spans the tested values of time lags (days). The RMSE values have units of MMcf (millions of cubic feet).

4.2 Time Evolution of Optimal Parameters

Another dimension to inspect within our optimization results is the time evolution of the optimal parameters. As an example tied to the results shown in the previous section, Figure 4 shows the time evolution of the optimal base temperature and time lags for CONUS from 1991 through 2005. The top panel shows the evolution of base temperature for HDD_w, and the bottom panel shows the evolution of time lag. Both panels have three different lines plotted atop one another: the connected blue circles show the raw outputs from our procedure for each year, the green line shows a smoothed representation of the line with blue circles, and the connected black diamonds depict the green line “snapped” to the nearest integer value for each year. The reason to include smoothed versions of the raw data is an acknowledgment that there seem to be meaningful trends within the evolution of the optimal parameters, and that the year-to-year jitter is perhaps consistent with the aforementioned broad basins associated with the global minimum each year. Based on these trends, it seems the nationwide, population-weighted optimal base temperature to use for HDDs has decreased from about 61 °F in the 1990s to about 58 °F more recently. Also, the nationwide optimal time lags appear to have held steady at about 9-to-10 days until about 2007 or so, after which there was a steady decrease to about 2-to-3 days recently.

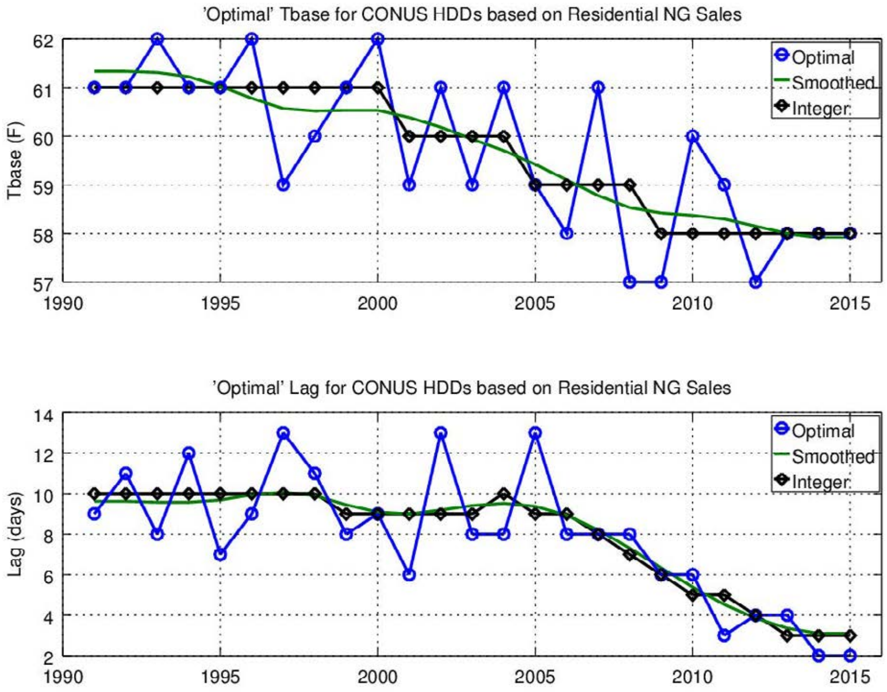

Figure 4: Time evolution charts of the optimal base temperatures (top) and time lags (bottom) to use for CONUS data. The blue circles show the raw values, the green lines show smoothed versions of the raw data, and the black diamonds show the nearest integer values to the smoothed values.

The blue circles in Figure 4 are the results from a simplified statistical procedure, so we do not want to over-interpret it. Nevertheless, it is encouraging that it is at least consistent with narratives we might have applied prior to undertaking this research project. Namely, regarding HDD_w, we would have expected the optimal base temperatures to decrease in time since HDDs themselves are a model relating indoor heating behavior to outdoor temperatures: newer housing units are presumably built with better insulation properties than older homes, thus allowing newer housing units to rely on space heating less intensely and on fewer days a year than older housing units. Regarding time lags, it is very interesting to see the steady decrease in effective time lag starting in 2007. According to the survey manager of EIA-857, the monthly natural gas sales survey whose results we have used in our analysis, the need for EIA to do in-house "calendarization" was discovered in late 2006. EIA warned survey respondents that a change would be coming soon, and apparently some began to attempt the calendarization themselves. Then, starting in 2010, EIA performed the calendarization itself, thus minimizing the time lag by construction.

4.3 Spatial Distribution of Optimal Parameters

Another dimension of interest when inspecting our optimal results is how different geographical regions compare spatially. Because we applied our procedure to all U.S. states within CONUS, we can actually make maps of the optimal parameters, colored by state, for any given calendar year. Figures 5 and 6 show such maps for the year 2015, again continuing the examples in the previous section. Keep in mind that the CONUS-wide optimal values for base temperature and time lag are 58 °F and 2 days, respectively.

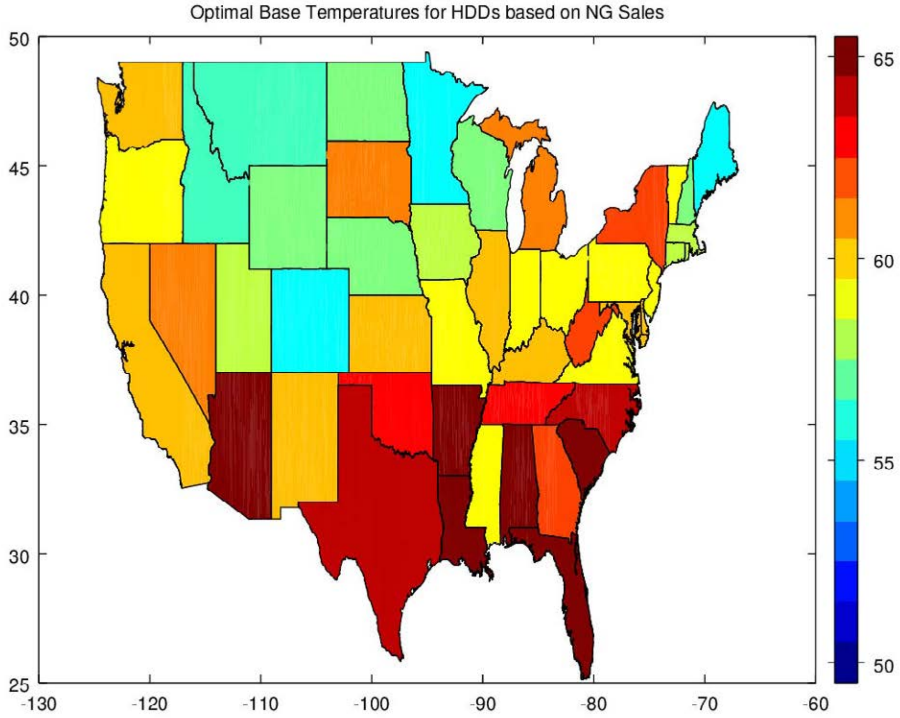

Figure 5: A map of optimal base temperature (°F) for HDD_w by U.S. state in 2015. The x-axis is longitude (degrees West), and the y-axis is latitude (degrees North).

Looking at the map in Figure 5, we might first note the varied texture to its coloring, with colors spanning from about 55 °F to 65 °F. Looking more closely, we can see that there are broader swathes of similarity, with the South mostly showing high optimal base temperatures, and the Northern Plains and New England showing mostly low optimal base temperatures. That there is roughly continuity to complement contiguity is a pleasing confirmation of the optimization process, since there is nothing within it to target this or ensure it. Further, the swathes we do see fit the narrative we might have offered before doing this analysis: housing units in colder regions are likely built with better insulation than housing units in states with milder winters.

Looking at the map in Figure 6, we might first note again the varied texture to its coloring, with most colors spanning from 0 to about 6 days, but there are two clear outliers with values of 9 and 13 days. Seeing that most U.S. states have very low lags is encouraging, given that we know that EIA has been doing its own in-house calendarization of the utility natural gas sales data since 2010. It appears that not all states have fully cooperated with the survey instructions, but many states clearly have, at least if we are to believe our simple optimization results.

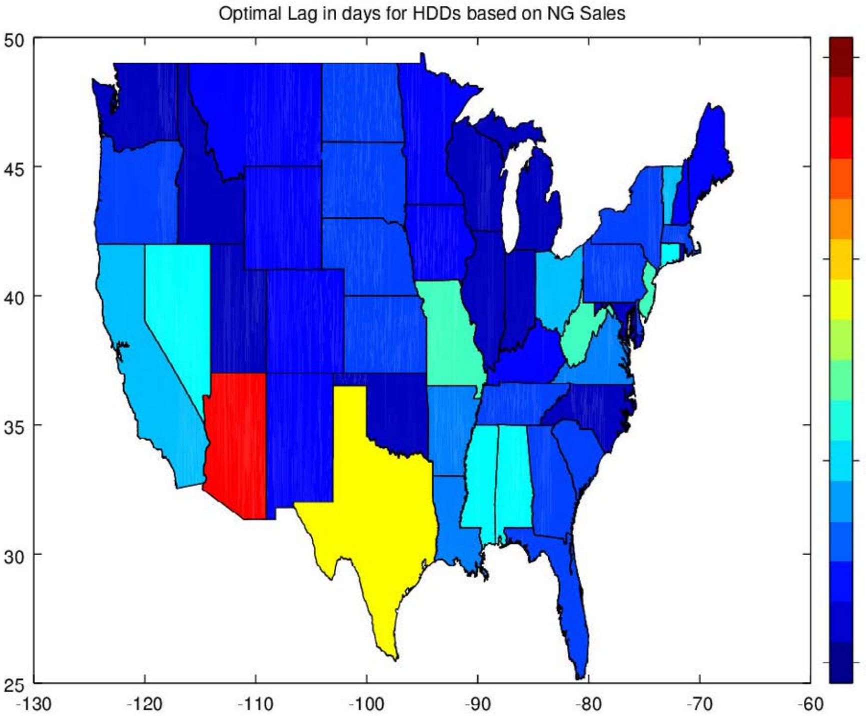

Figure 6: A map of optimal time lags (days) by U.S. state in 2015. The x-axis is longitude (degrees West), and the y-axis is latitude (degrees North).

4.4 Time Record of Space Heating Consumption

Here we return to the original motivation for our analysis, estimating the natural gas consumption attributable to space heating, which is dependent on HDDs. Having found the best values for base temperature and time lag to use for each calendar year, we can transform the daily HDD$_w$ data accordingly, aggregate into monthly values, and proceed with the simple linear regression as described in section 3.1. This will allow us to estimate the fraction of the total consumption attributable to space heating by $\sum_{i=1}^{12} \beta_1 \cdot HDD_{w,i}(B_{opt}, \tau_{opt})$, which we can then divide by the total consumption for that year to yield a percentage due to space heating. If we apply this to the CONUS data, we will obtain something close to the RECS estimates for natural gas consumption due to space heating. These estimates are not strictly comparable for several reasons, the two most important of which are: 1.) RECS only covers primary, occupied housing units (as opposed to vacation homes and vacant homes), whereas the EIA surveyed sales data covers the complete residential sector, and 2.) based on data availability constraints, we can only focus on CONUS in our analysis, whereas RECS includes HI, AK, and DC.

It is difficult to show definitely how these issues might affect the comparability of our estimates here to RECS, but we can speculate briefly. The fact that RECS only covers a portion of the full residential sector should lead to RECS estimating less overall total natural gas consumption than one finds in the complete sector, which just stands to reason because some fraction of the excluded housing units definitely consume natural gas. By similar reasoning, it should be the case that the RECS estimate for natural gas consumption due to space heating should also be a bit lower than the theoretical true value for the entire residential sector because some fraction of the excluded housing units will have consumed natural gas for space heating. Whether the excluded housing units use natural gas in the same proportion as all of the housing units included in RECS is an open question, but it seems prudent to compare RECS estimates for space heating to our own estimate here based on the *percentage* of total natural gas consumption used for space heating. Considering the fact that RECS includes HI, AK, and DC in its estimate, it is perhaps a worrying difference, but given that our approach is relying on population-weighted heating degree days, and given that the excluded states and DC have relatively low populations, we do not expect their exclusion to have too large an effect. Also, to the extent that it matters, the excluded states and DC do not share much in common as far as weather trends and expected HDDs, so perhaps there would have been some amount of "cancellation" by these regions had they been included in the first place. Either way, we proceed by comparing our estimates to previous RECS estimates based on percentages.

As an example of the kind of fit we obtain based on the optimal parameters, and how we do the decomposition into consumption due to space heating and all other consumption, we offer the case here in 1993, a year for which we also have a RECS estimate. Figure 7 shows a comparison of the reported natural gas sales data (blue circles) and our optimized simple regression model (red circles): the CONUS best-fit parameters for 1993 are 62 °F and 8 days. The bottom panel shows the same values as the red circles in the top panel, but here depicted as a bar graph. The bars have been decomposed into a constant "base load" component, colored in gray, and a component due to space heating, colored in red. Summing the total consumption within the red bars, we find 3,480 BCF (billions of cubic feet of natural gas). Summing the total annual consumption of natural gas in 1993, we find 4,925 BCF. This yields an estimate of 70.6% for the percentage of total consumption used for space heating. This can then be compared to the RECS estimate from that year, which is 69.6%. Given the assumptions and uncertainties involved both in RECS and our simple analysis here, we can likely consider these statistically close, if not equivalent.

Figure 8 shows a time series over the extent of our analysis period comparing our year-by-year estimates to the available RECS estimates over the same period. Our "external-to-RECS" estimates are plotted as blue circles, and further emphasized as heavy blue squares in years where there are also RECS estimates available. The first year of comparison in our analysis period is 1993, which is the example shown in Figure 7. We can see in this figure that our own estimates for the percentage of natural gas consumption due to space heating has varied from as low as about 68% to nearly 74%. Comparing our numbers to RECS, we see there is good agreement in 1993, 1997, and 2001; however, there is stark disagreement in 2005 and 2009, though we note that the y-axis of this figure has been greatly zoomed-in, so the differences are not so large in an absolute sense. We are at a loss to explain this apparent disagreement, though we have alerted the RECS team to this discrepancy and they are looking into it. We will note that this finding was an unintended by-product of our research: we had no reason to question any historical RECS results. Rather, we were motivated to begin this research knowing that RECS 2015 end-use estimate results will be published soon, and we were looking for an external estimate to "sanity check" those estimates before public release.

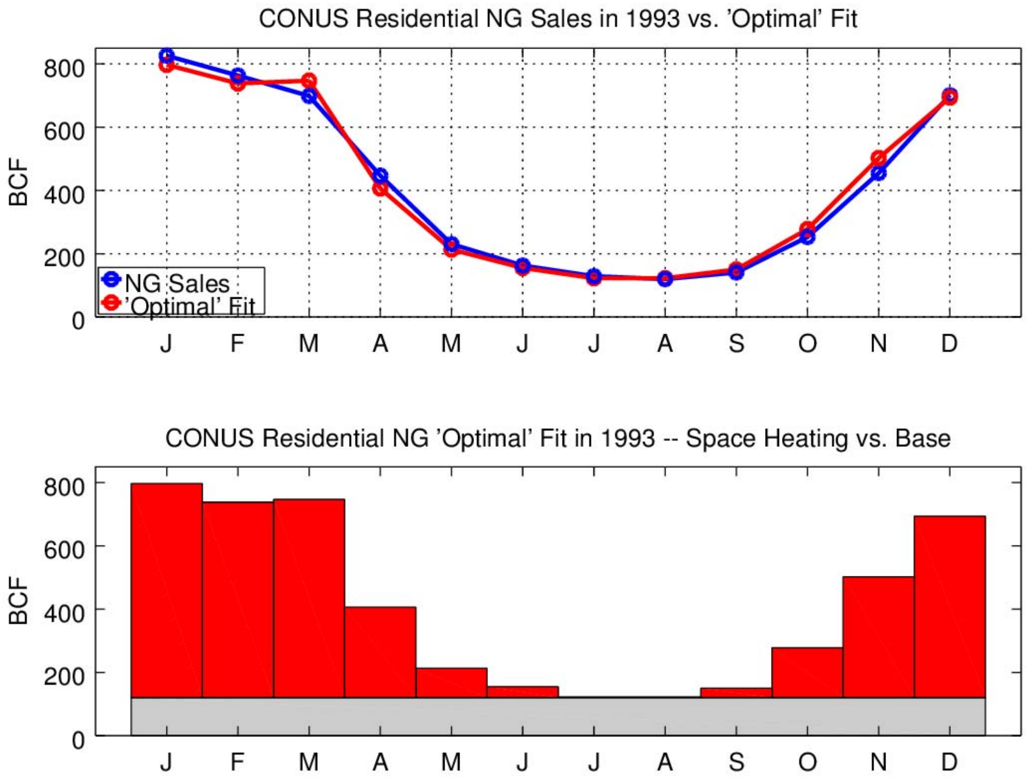

Figure 7: An example of our simple linear regression analysis for how much natural gas was consumed within CONUS in 1993 for space heating purposes. The top panel compares the reported monthly utility totals of natural gas sales to our optimized simple linear regression prediction. The bottom panel shows how we decompose our model prediction into a constant "base load" (gray) and a component due to space heating (red).

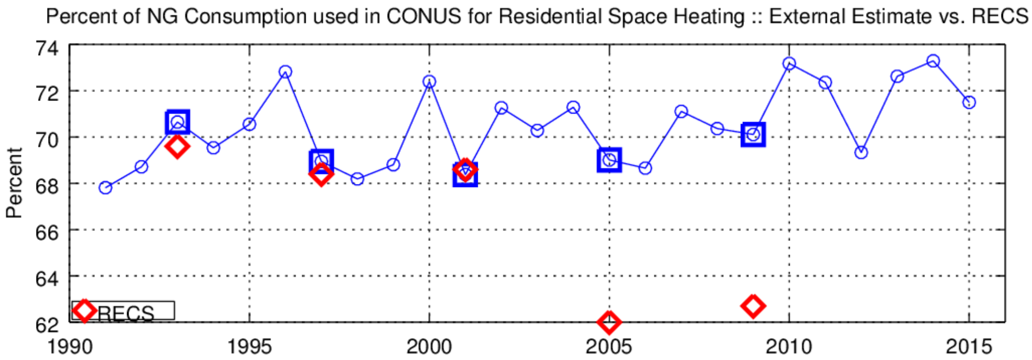

Figure 8: A time series of our estimates for the percentage of natural gas consumption due to space heating within CONUS in each calendar year over our analysis period, compared to the available RECS estimates within that time period.

5. Conclusions

To summarize, we have sought to find a method for estimating the energy consumption for seasonal end uses like space heating and air conditioning. We want the method to be completely external to RECS so that it could be potentially used as verification or a sanity check. We have proposed using an approach of brute-force optimization atop of simple linear regression to best isolate the portion of monthly consumption attributable to weather variations, as encoded in population-weighted degree days. The optimization is used to estimate the best values to use for two sensitive degrees-of-freedom: the base temperature with respect to which the population-weighted degree days are calculated, and the unknown time lags that exist in utility-reported consumption data. We find that our method gives reasonable estimates for these "optimal" parameters: both their time evolution and spatial distribution are consistent with narratives we could have spun before this analysis was undertaken. The final resulting estimates for the percentage of natural gas consumption attributable to space heating seem reasonable in their own right. Further, the estimates agree with earlier estimates from RECS, though we find a stark disagreement in recent years, which we cannot explain as of yet.

There are of course many paths forward in this line of research, including revisiting some of the decisions we have made to get to this point. One obvious next step would be to compare the national estimates we have obtained working directly with CONUS data to the estimates we would obtain from separately adding up the estimates from each individual U.S. state. This latter approach would have the benefit of allowing each state to have its own optimal base temperature and time lag, which would probably be more accurate than assuming all states share the same optimal values. Also, upon further reflection, we see now that we could have extended our range of prior base temperatures to somewhat higher values than 65 °F. This seems justified in retrospect knowing that we found some U.S. states have their optimal B values as 65 °F, and it is generally a good idea to search further if one's optimization problem ends up choosing a value that is an endpoint in one's search domain. Another next step would be to somehow acknowledge the fact that water heating is another end use for natural gas that also exhibits seasonality, though not necessarily in direct proportion to HDDs as space heating is thought to respond. This means that the "base load" in natural gas consumption is not necessarily flat as we have depicted it in gray in Figure 7. And of course, the most important next step is to attack electricity sales/consumption data and make air conditioning estimates; however, there is also electric space heating to worry about, so there are potentially four degrees-of-freedom to control for instead of just two as we encountered with natural gas and space heating only.

Acknowledgements

I would like to gratefully acknowledge support from my supervisor, Hiroaki Minato. He has both supported this work in the literal sense of allowing it to proceed, and he has also been a resource with whom I can talk over issues. I would also like to acknowledge the helpful feedback I have received from ASA's Subcommittee on Energy Statistics, who convene twice a year at EIA and provide feedback on ongoing research projects. Mimi Goldberg has been a particular help from this committee.

References

Climate Prediction Center, National Oceanic and Atmospheric Administration, 2017. ftp://ftp.cpc.ncep.noaa.gov/htdocs/degree_days/weighted/daily_data/

U.S. Energy Information Administration, 2017. *Monthly Energy Review*, Table 2.2. https://www.eia.gov/totalenergy/data/monthly/#consumption

U.S. Energy Information Administration, 2017. *Residential Energy Consumption Survey*. https://www.eia.gov/consumption/residential/

Fels, M. F., 1986. "PRISM: an introduction." *Energy and Buildings*, **v.9** (1-2), pp. 5-18.

Fels, M. F., and M. L. Goldberg, 1986. "Using the scorekeeping approach to monitor aggregate energy conservation." *Energy and Buildings*, **v.9** (1-2), pp. 161-168.

Goldberg, M. L., 1982. *A Geometrical Approach to Nondifferentiable Regression Models as Related to Methods for Assessing Residential Energy Conservation*, Ph.D. thesis, Department of Statistics, Princeton University.