ECON_1 TASK

ANDREW BLOHM

1. Abstract

In this paper we estimate market shares for power production technologies using a multinomial regression framework. The contributions of this paper are two fold: first, the data cleaning process required for the subsequent analysis is significant and not likely to have been accomplished before. We make this data available online at `https://github.com/andymd26/fuzzy-waffle`. Second, GCAM is an integrated assessment model, which presently uses historic data to parameterize the electricity production technology market shares, which are then nonvarying. This work incorporates important determinants in the relationship between preferences, market factors, and market share into the GCAM framework and thus strengthens its output.

2. Theory: Discrete choice models

An agent n faces a choice of technology alternatives j. We assume that each agent is attempting to maximize their own utility U_{nj} through the selection of a technology alternative. However, U_{nj} is unobservable and consequently unknown to the researcher. Instead, we can observe information about the available choice set (x_{nj}), as well as the agent (s_n). Using this information we can specify a representative utility function, which is $V_{nj} = V(x_{nj}, s_n)$.[1]

We can write each agents utility function as a function of observed and unobserved utility, $U_{nj} = V_{nj} + \epsilon_{nj}$, where the distribution of the error term (ϵ_{nj}) depends on the researchers specification of the observed utility (V_{nj}). Given that we cannot observe ϵ_{nj} directly we instead derive the probability of a technology choice. The probability that the agent chooses technology j (i.e., the expected value of the indicator function) is $Pr(j|x) = Pr(I[h(x,\epsilon) = j] = 1) = \int I[h(x,\epsilon) = j] \cdot f(\epsilon) d\epsilon$.[2]

The probability of an individual choosing the alternative that they actually chose is: $\prod_i (P_{ni})^{y_{ni}}$, where y_{ni} is either one for the selected alternative or zero for the remaining unselected options. Given that the exponents are either zero or one the result reduces to the term P_{ni}, which is just the probability of the chosen alternative. The likelihood function then is $\mathcal{L}(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}}$, which can be rewritten as $ln(\mathcal{L}(\beta)) = \sum_n \sum_i y_{ni} ln\,(P_{ni})$, where the specification of P_{ni} depends on the specification of the model.

3. Data

The dependent variable of our analysis is the annual summer capacity addition by prime mover and primary fuel type weighted by the associated capacity factor. Summer capacity represents

Date: September 21, 2016.

[1] $U_{nj} \neq U_{nj}$ given the unobservable aspects of the utility function.

[2] The unobserved parameters ϵ_{nj} follow a probability density distribution $f(\epsilon)$, which in the logistic framework we assume to be distributed extreme value with variance of $\frac{\pi^2}{6}$.

the maximum capacity as determined by testing during the summer peak demand period (i.e., June 1st to September 30th) and is lower than the nameplate capacity since it includes electricity consumption by auxiliaries and the station itself ()(EIA Glossary, 2016).

We generate the variable using information from the EIA Form 860, which is a survey of electric utilities and contains detailed information down to the generator level. The variable is calculated for all unique pairings of primary fuel and prime mover technology using the in-service date provided in the survey.[3][4] Only a small number (~ 50) of the approximately 375,000 entries in the data are missing the in service date. However, one potential issue with using this approach is that it does not account for fuel switching behavior or derating in the data. However, given the purpose of this exercise this should not bias the results.

In Figure 1 we show how the unweighted annual summer capacity additions by electricity production technology and primary fuel is changing over time. Over the past decade there has been significant investment in natural gas fired resources and renewable resources with little investment in coal fired resources.

The capacity factor weighting is calculated using the annual survey results for Form EIA-923 (), which surveys approximately 4,100 power plants each year.[5] The form contains plant level data (i.e., one level higher than generator) on technology choice, environmental compliance, net generation, etc. We use the data to calculate the average capacity factor for each primary fuel k and production technology j pairing, which is

$$\text{Avg. CF} = \frac{1}{n_{jk}} \sum_{j} k \frac{\text{Annual Net Generation (MWh)}}{\text{Size (MW)} * 8760(hrs)} \tag{1}$$

Where, $n_j k$ is the number of plants with primary fuel k and production technology j. Unfortunately, the size of the plant is not a variable captured by Form EIA-923. Instead, we identified the size using data from Form 860 and merged the two sources together. This process has the potential to introduce more error into our calculations, as Form EIA-860 and Form EIA-923 disagree on the number of generators active at each site. The advantage of this approach is that the capacity factors are tailored to the primary fuel and production technology, as opposed to capacity factors found in the literature, which are solely based on production technology and tend to be available for a short amount of time.

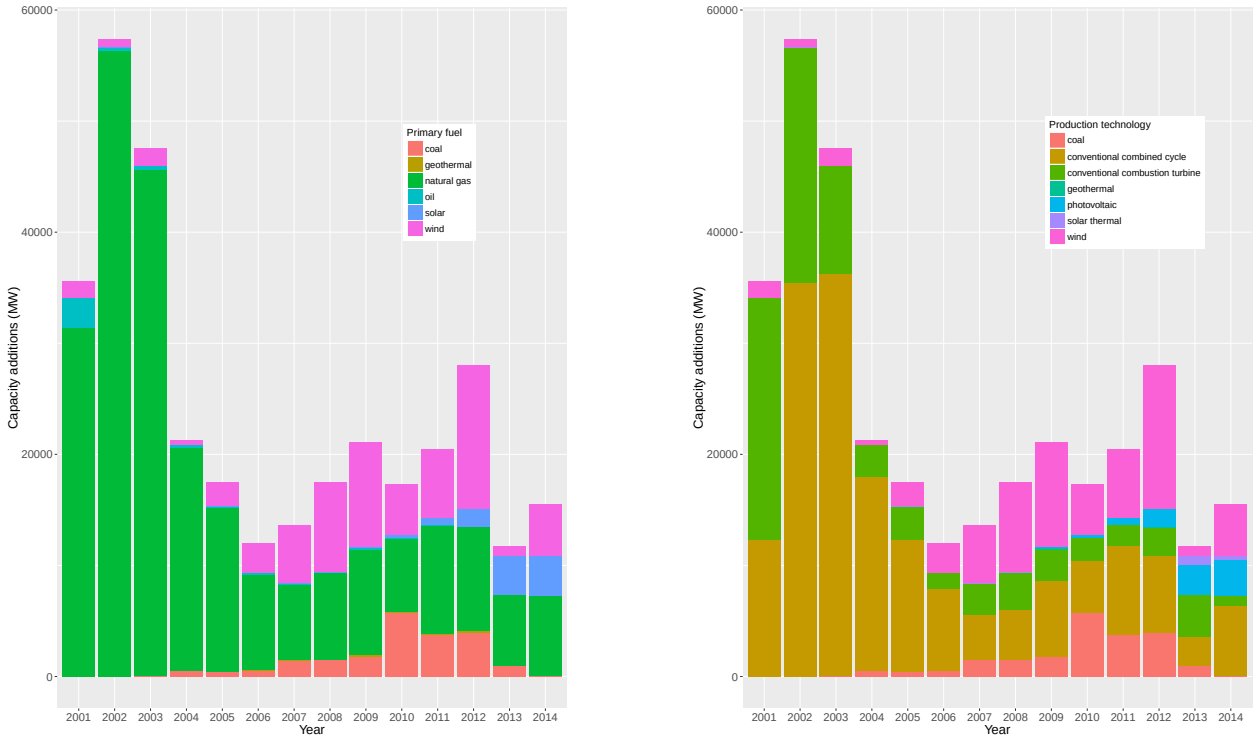
From the Annual Energy Outlook (AEO)series, specifically the underlying assumptions on characteristics of new and existing power plants, we gather cost estimates for various power production technologies.[6] Each annual energy outlook estimates the characteristics of new generating plants, including operational characteristics and cost assumptions. Each year the AEO estimates project

[3]Fuel switching, plant derating and other issues lead to negative numbers if we try and calculate the capacity additions using the change in summer installed capacity even after accounting for plant retirements.

[4]A survey of the data appears to show that this could be an issue of plant derating. For example, Utility 13781 operates Plant 3982, Generator 4, which has an installed summer capacity of 22 MW and a steam turbine to convert biomass to electricity. The unit exists in the database for the period 1990 to 1995 but in 1992 the installed capacity drops to 20 MW. By 1995 the capacity returns to 22 MW before the unit then drops out of the survey for the period 1996 to 2004. When it returns in 2004 it has a capacity of 22.6, which declines to 22 MW in 2005. By 2010, the installed summer capacity is 15 MW, which remains its capacity to present. During this period there is no change in the operating status of the unit nor was it retired and brought back online.

[5]Further information on the method can be found in the Supplemental Material.

[6]Unfortunately, the types of power plants listed in the AEO do not directly map onto the information we have in Form 860. As a result, we created our own mappings that can be found in the supplemental material.



(A) Capacity addtion by primary fuel, 2001-2014

(B) Capacity addition by primary mover, 2001-2014

FIGURE 1. Capacity additions in the electricity production industry by primary fuel and primary mover (EIA 2015)

lead time, average size, heatrate, fixed and variable operation and maintenance costs, and overnight costs for a suite of first and n^{th} of its kind power production technologies.[7]

We use the overnight capital cost assumptions from each AEO to approximate the investment cost for each technology in each year of the analysis. The earliest AEO for which the overnight cost is available is the 1997 AEO; having been published in each subsequent year. However, as new technologies emerge and existing technologies evolve the categories within the overnight cost database have necessarily changed. Therefore, the database does not have consistent categories over time. In our model we use the overnight cost, as well as the fixed and variable operations and maintenance costs. We considered using the estimates of the heat rate from the AEO but ended up using other sources, which we will discuss later.

The overnight cost is the cost of a construction project with no interest incurred. We weight the overnight cost by the reciprocal of the capacity factor in order to standardize the data across

[7]For mature technologies we would expect the first and n^{th} cost estimate to be very similar, however, we would not expect this for emerging and immature technology.

production technologies.[8] Fixed operations and maintenance costs are expenses that don't vary with production such as routine preventive and predictive maintenance, general and administrative expenses, fees required to participate in NERC and other regulatory bodies, etc. (EIA 2010). Variable operation and maintenance costs are expenses that by definition vary with production such as, water, disposal expenses, power purchases, consumable materials and supplies, etc. (EIA 2010). We adjust the overnight cost

We assume that fuel costs are the predominant operating cost for fossil fuel generation plants and are thus, an important factor in power plant production technology selection. Fuel price data for the period (1/1986-5/2016) is from the Energy Information Administration (EIA) sources. The Energy Information Administration (EIA) in the Electric Power Monthly report publishes electric utility receipts of and average cost for fossil fuels used in the power generation sector. We generate a monthly time-series dataset for the period 1/1986-5/2016 using the Electric Power Monthly reports from June 1996 (Table 26), January 2010 (Table 4.2), and May 2016 (Table 4.2). The data includes price information for coal, petroleum liquids, and natural gas inputs. The prices listed are averages and in nominal units (i.e., unadjusted for inflation). For continuity purposes we use the total petroleum price from the June 1996 report instead of the heavy oil price.

We adjust the fuel price for the efficiency of each production technology using the heatrate. The heatrate is a measure of efficiency in converting fuel into electricity. It is defined as the amount of fuel necessary to generate one kWh of electricity. We calculate the average annual heat rate using the Form EIA-860 and the California Energy Almanac. For more information on the method please see the Supplemental Material.

4. Model

The probability of choosing technology choice i is equal to $P_{ni} = P(V_{ni}+\epsilon_{ni} > V_{nj}+\epsilon_{nj} \quad \forall \quad j \neq i)$. This can be rewritten as the following $P_{ni} = (\epsilon_{nj} < \epsilon_{ni} + V_{ni} - V_{nj} \quad \forall \quad j \neq i)$. Given that the density of the unobserved utility is $f(\epsilon_{nj}) = e^{-\epsilon_{nj}} e^{-e^{-\epsilon_{nj}}}$ and the cumulative distribution function is $F(\epsilon_{nj}) = e^{-e^{-\epsilon_{nj}}}$ then the probability of choosing technology i is $Pr_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-(\epsilon_{nj}+V_{ni}-V_{nj})}} \right) e^{\epsilon_{ni}} e^{-e^{\epsilon_{ni}}} \, d\epsilon_{ni}$. [9] The integral simplifies to $P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}$.[10][11][12]

[8]The transformed data now reflects the investment needed to achieve a standardized summer installed capacity of 1 MW.

[9]The Pr_{ni} can be interpreted as the joint probability of not choosing alternative i multiplied by the probability of choosing alternative i.

[10]V_{ni} is usually specified to be linear in parameters (i.e., not multiplicative).

[11]The multinomial logit model estimates $J-1$ equations that compare each of the $J-1$ categories to the baseline category J (i.e., the multinomial logit model reduces to the standard logistic regression in the case of $J = 2$.

[12]The multinomial and Poisson models have an equivalence that we can exploit in order to model the market shares using a generalized linear model (rather than the multinomial logit framework). One advantage of this is our own familiarity with the GLM package in R, as compared to the mlogit package.

If we have a set of Poisson distributed random variables, such as the annual capacity additions in the power sector, $X_1 \sim P(\lambda_1)$, $X_2 \sim P(\lambda_2)$, ..., $X_k \sim P(\lambda_k)$ where $\lambda_1 \neq \lambda_2 \neq \ldots \neq \lambda_k$ (at least not necessarily equal) then the conditional distribution (i.e., probability distribution for a sub-population) of the random variables $X = (X_1, X_2, \ldots, x_k)$ is multinomial, $X \sim Mult(n, \pi)$, where $n = \sum_k X_k$ and $\pi_j = \frac{\lambda_j}{\sum_k \lambda_k}$.

The relationship between the Poisson and multinomial model allows us to estimate a Poisson model using a generalized linear model and then calculate the log-odds from the resulting coefficients. If we treat the random counts (Y_{ij}) (i.e., capacity additions by production technology and primary fuel) as Poisson random variables with mean (μ_{ij}) then we can estimate the following log-linear model.

We can normalize the model for the scale of the utility by setting the variance equal to $\frac{\pi^2}{6}$. the scale of the utility is irrelevant to the alternative chosen needs to be normalized as the alternative with the highest utility is the same no matter the scale (i.e., the alternative chosen is the same in these two formulations, $U^0_{nj} = V_{nj} + \epsilon_{nj}$ and $U^1_{nj} = \lambda V_{nj} + \lambda\epsilon_{nj}$). Researchers usually do this through the normalization of the variance of the error terms. If we assume the error terms are iid then we can implement this by dividing the model coefficients by the standard deviation of the unobserved portion of utility. The coefficients now represent the 'effect of the observed variables relative to the standard deviation of the unobserved factors' (Train 2009, p. 24). The error variances in a traditional logit are normalized to $\frac{\pi^2}{6}$ (Train 2009).[13]

The absolute value of U_{nj} does not matter only the difference in utility between alternatives since $Pr(U_{ni} - U_{nj} > 0, \quad \forall \quad j \neq i)$. This fact has implications on variable selection, in that we only include parameters that capture differences across the alternatives. The absolute value of the constant term (k_j) does not matter only the difference. To do this we need to normalize the absolute value of one constant by setting it equal to zero. The remaining $J-1$ constants are then the difference in constant values for each alternative, as compared to the baseline.[14] The same holds true for any variable that doesn't vary for an individual between alternatives (i.e., individual income). This issue can be addressed through the normalization of the variable or by interacting it with attributes of the alternatives.

The multinomial logit model imposes the restriction that the error distribution is independent and identical over alternatives (i.e., technology choice selected are iid extreme value (i.e., unobserved portion of the utility is uncorrelated over the choice set and have the same variance) (Wen and Koppelman 2001).[15] Should we believe correlation exists across alternatives then we either respecify the model to account for the correlation structure; explore other alternatives later that do not have this restrictive assumption (i.e., GEV, probit, and mixed logit); or continue with the logit specification as an approximation.

We assume that correlation exists between some technology alternatives, which necessarily violates the independence of irrelevant alternatives, which assumes that for two alternatives *coal* and *ng*, we have $P_{coal} = \frac{e^{V_{coal}}}{\sum_j e^V_j}$ and $P_{ng} = \frac{e^{V_{ng}}}{\sum_j e^V_j}$, which means that the ratio of the two is $\frac{P_{coal}}{P_ng} = \frac{e^{V_{coal}}}{e^{V_{ng}}}$. The ratio is only a function of characteristics of each alternative and not related to other alternatives. However, the choice between some options might violate this assumption, in that the error terms are correlated for alternatives in the same 'nest'. The generalized extreme value model (GEV) allows for correlation in the unobserved factors over alternatives (if the correlation amongst these factors is zero then the GEV is equivalent to the logit model). To estimate this model requires the nested logit model, which allows for error terms to be correlated within nests and uncorrelated between nests (McFadden 1978).

The Generalized Extreme Value model (GEV) also known as the nested logit model assumes that the error terms are distributed, whereby in Equation 2 if $\lambda_m = 1$ then we have no correlation and the model is equivalent to the multinomial logit discussed previously.

---

[13] If we feel that we have heteroskedastic errors at any point then we need to review the discussion on p25 (basically just normalize one of the subpopulations and then the variance estimates for the other groups will be relative to that baseline).

[14] This can be implemented in the mlogit R package through the 'reflevel' command.

[15] It should be noted that through proper model specification we could reduce the importance of the independence assumption. If we specify the utility model well (i.e., $U_{nj} = V_{nj} + \epsilon_{nj}$), accounting for the important differences between alternatives, then the error term should be white noise (i.e., the error for one alternative provides little to no information about the error for another alternative).

$$exp\Big(- \sum_{m=1}^{M} \Big(\sum_{j \in B_m} e^{\frac{-\epsilon_j}{\lambda_m}} \Big)^{\lambda_m} \Big) \tag{2}$$

It can be shown then that the probability of choosing option j in nest l is

$$P_j = P(j|l)P(l) = \frac{e^{Z_j/\lambda_l}}{\sum_{k \in B_l} e^{Z_k/\lambda_l}} \cdot \frac{e^{W_l + \lambda_l I_l}}{\sum_{m=l}^{M} e^{W_m + \lambda_m I_m}} \tag{3}$$

whereby is generally referred to as the inclusive utility, which is calculated as the $I_l = ln\Big(\sum_{k \in B_l} e^{Z_k/\lambda_l} \Big)$ and is the link between the upper and lower model. The first term of the nested logit structure represents the lower level of the model and is the conditional probability that an alternative is selected given that it is in that nest. The second term represents the expected utility of the nest m () mlogit docs. We implement this nested logit structure using the mlogit package in R (Croissant 2013)

The determinants of the choice made by each individual amongst the entire choice set can be classified according to three indexes: the properties of each alternative, the choice situation, and properties of the individual (Croissant mlogit package). At least three types of variables are possible in the multinomial logit framework: alternative specific variables (x_{ij}) with generic effects (β); individual specific variables (z_i) with an alternative specific coefficient (γ_j); and alternative specific variables (w_j) with alternative specific coefficients (λ_j)(Croissant 2013).

We consider the choice situation of our data (i.e., each year), as a repeated cross section. In that we aren't accounting for any relationships (at least at this point) between the choices made in one year and any subsequent years. In our case, we have information on each of the alternatives (i.e., cost, operational characteristics, etc.) but we consider the data as a repeated cross section. For the analysis we use the mlogit package in R for the data analysis (Croissant 2013).

For a discrete choice model the choice set must be mutually exclusive, exhaustive, and finite from the decision makers point of view. In our particular circumstance, we can redefine the choice set, as it is possible to use more than one technology, to be any combination of the technology choices such that the set is then mutually exclusive (i.e., A only, B only, or both A and B). To ensure that the list is exhaustive might require that we include a none of the above option as well. An underlying assumption of discrete choice analysis is that the selection of a decision alternative is necessarily mutually exclusive (i.e., a decision maker may select one and only one alternative per choice situation). However, the construction of our dependent variable is naturally at odds with this formulation since it does not preclude the selection of more than one technology (i.e., total capacity additions by primary mover and primary fuel). To address this issue, we create an artificial choice situation whereby we assume a decision maker is making a decision for each 1 MW block of installed summer capacity in that year.[16][17] In this way, we achieve mutually exclusive decisions that account for differences in total investment between years. The drawbacks of the approach include, smaller standard errors than warranted, and a failure to account for the average size of each decision alternative (i.e., we violate the assumption of independence that exists between choice situations).

[16]For example, if in Year 1 there were capacity additions of 5 MW and 10 MW, respectively for coal and natural gas then we would reformulate the data set as 5 decisions where coal was selected over natural gas and 10 situations where the alternative held true.

[17]We can't use the percentage allocated to each technology in each year because the total investment between years varies.

$$log\Big(\frac{u_{ij}}{u_{ik}}\Big) = (\alpha_j - \alpha_k) + x_{ij}^T(\beta_j - \beta_k) \tag{4}$$

This is the odds that the observation i would be in category j relative to the category k. The parameters of the multinomial model then are $a_j = \alpha_j - \alpha_k$ and $b_j = \beta_j - \beta_k$. This is equivalent to the multinomial logit model, which assumes a linear model for the log-odds of each response (see equation 5).

$$\gamma_{ij} = log\Big(\frac{u_{ij}}{u_{ik}}\Big) = \alpha_j + x_i^T\beta_j \tag{5}$$

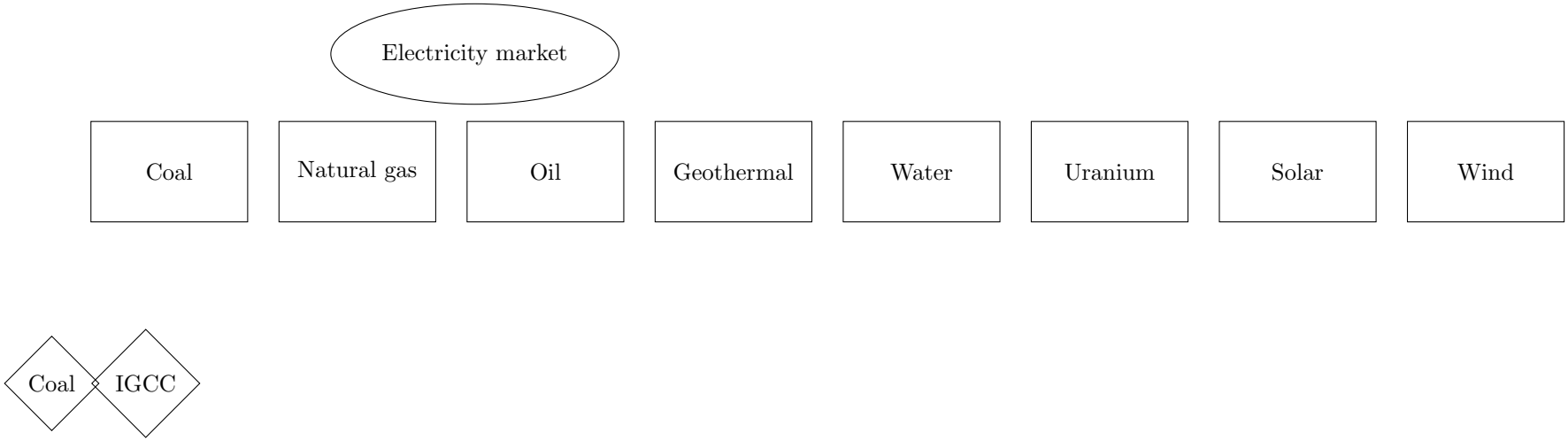
the relationship between the logit probability and representative utility is S-shaped which means that if representative utility is very low (relatively), a small increase in the utility of the alternative has little effect on the probability of its being chosen (from a policy perspective this would be like choosing between investing in an area with poor bus service or an area with sufficient bus service: more impact in the area with sufficient service on bus ridership)

5. Results

6. Discussion

7. Next steps

We might consider the nested logit models to incorporate the specific technology choice once the fuel choice has been made. glm implementation with percentage dependent variable and a weight equal to the change in installed MW because the percentages are not equivalent between the years (i.e., the absolute value is different than the relative percentage).


Electricity market
Coal
Natural gas
Oil
Geothermal
Water
Uranium
Solar
Wind
Coal
IGCC

References

Croissant, Yves (2013). *mlogit: multinomial logit model*. R package version 0.2-4. URL: https://CRAN.R-project.org/package=mlogit.

EIA (2010). *Updated Capital Cost Estimates for Electricity Generation Plants*. Tech. rep. U.S. Energy Information Administration.

– (2015). *Form EIA-860 detailed data*. URL: https://www.eia.gov/electricity/data/eia860/.

McFadden, D. (1978). "Spatial interaction theory and planning models". In: *Modeling the choice of residential location*. Ed. by A. Karlqvist. North-Holland, Amsterdam, pp. 75–96.

Train, Kenneth E (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Wen, Chieh-Hua and Frank S Koppelman (2001). "The generalized nested logit model". In: *Transportation Research Part B: Methodological* 35.7, pp. 627–641.