ECON_1 TASK

ANDREW BLOHM

1. Abstract

In this paper we estimate market shares for power production technologies using discrete choice analysis tools. 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, it proposes improved share weights for the GCAM electricity sector model, which are presently 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, j = 1, . . . , J. We assume that each agent is attempting to maximize their own utility U_{nj} through the selection of a technology alternative (i.e., each agent chooses technology i iff $U_{ni} > U_{nj}\ \forall\ i \neq j$).[1] However, U_{nj} is unobservable and consequently unknown to the researcher. Instead, we can observe information about the available choice set ($x_{nj}\ \forall\ j$), as well as the agent ($s_n\ \forall\ n$). Using this information we can specify a representative utility function, which is $V_{nj} = V(x_{nj}, s_n)$ (Train 2009).[2]

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 n chooses technology j is

$$\begin{aligned} &Pr(U_{ni} > U_{nj} \forall i \neq j) \\ &Pr(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj} \forall j \neq i) \end{aligned} \tag{1}$$

This is a cumulative distribution (i.e., the probability that the difference in the error terms is less than the difference in observed utility). We can rewrite this as the following, where I is an indicator function in the event that the condition in the parentheses is true and zero otherwise.

$$Pr(j|x) = \int I(\epsilon_{nj} - \epsilon_{ni} < V_{ni} - V_{nj}) \cdot f(\epsilon_n) d\epsilon_n \tag{2}$$

The specification of the density ($f(\epsilon_n)$) is what differentiates different discrete choice models from one another (i.e., assume varying density distributional forms) (Train 2009). In a later section we discuss the specification of the density that we use.

Date: September 30, 2016.

[1] 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.

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

3. Data

In this section we introduce the data used in the multinomial regression model detailed in the following section. For further information on any of the data cleaning methods or underlying assumptions made please see the supplementary materials.

The dependent variable of our analysis is the weighted annual summer capacity addition for each unique prime mover and primary fuel type pairing.[3] The installed summer capacity represents 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 2016c). We consider the annual capacity additions as a repeated cross section rather than as panel data because while a time dimension exists, it is not the case that we are following the decision making of individual actors. Instead, we only see the aggregated result of individual decision making for each year.

For a discrete choice model, the choice set must be mutually exclusive, exhaustive, and finite from the decision makers point of view. However, the construction of our dependent variable is at odds with these assumptions given that capacity additions are made to a variety of technologies each year. 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.[4] In this way, we achieve mutually exclusive decisions that account for differences in total investment between years.[5] However, the approach has several drawbacks including, smaller standard errors than warranted because we are artificially increasing the number of choices undertaken, 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).

We generate the dependent variable using information from the EIA Form 860, which is a survey of electric utilities and contains detailed information down to the generator level (EIA 2016a).[6] For all unique pairings of primary fuel and prime mover technology we use the in-service date (provided in the survey) to generate the annual capacity additions.[7] Only a small number (~ 50) of the approximately 375,000 entries in the data are missing their in-service date. However, one potential issue with this approach (as discussed in the footnote) is that it does not account for fuel switching behavior or derating in the data. However, given the initial purpose of this exercise it should not bias the results though additional analyses might need to take this behavior into account.

---

[3]We weight the annual installed capacity by the associated capacity factor of each prime mover and primary fuel type pairing.

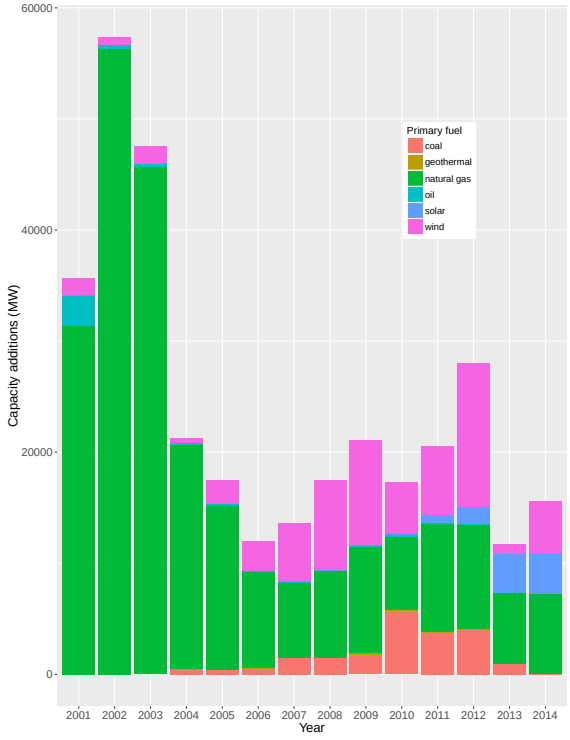
[4]For example, if in Year 1 there were capacity additions of 5 MW and 10 MW, respectively for coal and natural gas then we would reformulate the data set as 5 decisions where coal was selected over natural gas and 10 situations where the alternative held true.

[5]Another approach we considered was modeling the market shares of each technology each year. However this approach does not account for the variation in total capacity investment across years.

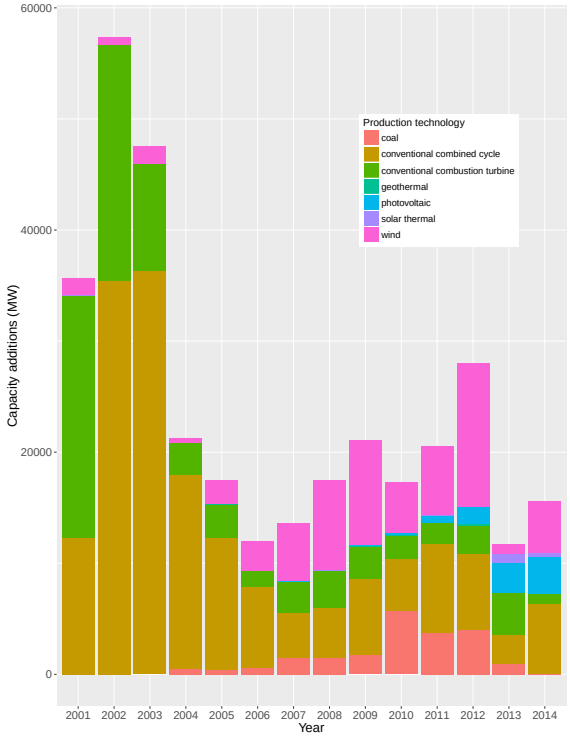
[6]There are three levels in the survey: utility, plant, and generator. Utilities can own more than one plant and there can be more than one generator at each plant.

[7]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 between years (even after accounting for plant retirements). 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.

In Figure 1, we show that the unweighted annual summer capacity additions by electricity production technology and primary fuel is changing over time.[8] From the figure, we can see that over the past decade there has been significant investment in natural gas fired resources and renewable resources with little investment in coal fired resources. Also apparent in the data is a complete lack of investment in integrated gasification combined cycle and nuclear plants. Given that none of these plants has been selected over the period of analysis they will continue to not be selected by the multinomial logistic model (at least the current iteration).



(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 2016a)

The capacity factor used to weight capacity additions for each prime mover, primary fuel pairing is calculated using the annual survey results for Form EIA-923, which surveys approximately 4,100 power plants each year (EIA 2016b).[9] 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

[8] By unweighted we mean that the installed capacities are not adjusted by the capacity factor.

[9] Further information on the methods we used can be found in the Supplemental Material.

to calculate the average plant capacity factor for each primary fuel k and production technology j pairing, which is

$$\text{Avg. CF} = \frac{1}{n_{jk}} \sum_j \sum_k \frac{\text{Annual Net Generation (MWh)}}{\text{Size (MW)} * 8760(hrs)} \tag{3}$$

Where, n_{jk} is the number of plants with primary fuel k and production technology j. Unfortunately, the size of each plant is not a variable captured by Form EIA-923. Instead, we identify the size pf each plant using data from EIA Form 860 and merge the two data sources using the unique plant identifier. This process introduces error into our calculations, as Form EIA-860 and Form EIA-923 do not always agree on the number of generators active at each site during a particular period. However, 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 (usually ∼2010 to present).

From the Annual Energy Outlook (AEO), specifically the underlying assumptions on characteristics of new and existing power plants, we gather cost estimates for various power production technologies. Each annual energy outlook estimates the characteristics of new generating plants, including operational characteristics and cost assumptions. Each year the AEO estimates project 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.[10] Unfortunately, the types of power plants listed in the AEO supplemental materials do not directly map onto the information found in Form 860. As a result, it was necessary to map the AEO cost estimates onto the information found in EIA Form 860. Details on the mapping exercise can be found in the supplemental material.

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 overnight cost is the cost of a construction project with no interest incurred. 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.

We weight the overnight cost by the reciprocal of the capacity factor in order to standardize the data across production technologies.[11] From the AEO supplemental material we also select the fixed and variable operations and maintenance costs for each technology. 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 considered using the estimates of the heat rate from the AEO documentation but ended up using other methods, which are discussed later.

The literature suggests 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

[10]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.

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

data for the period (1/1986-5/2016) is available from 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. Combining the fuel price and heat rate then generates a new measure that is the fuel cost (USD) necessary to generate 1 MWh of production for each unique prime mover and primary fuel pairing. We calculate the average annual heat rate using the Form EIA-860 and the California Energy Almanac (California Energy Commission 2016). For more information on the method please see the Supplemental Material.

4. Model

In this section we discuss the implementation of the model. We implement both a multinomial logit model and a nested multinomial logit model to estimate the parameter values of interest. We compare the nested and unnested models to test the independence from irrelevant alternatives (IIA) assumption. The nested model parameters are then compared with the GCAM electricity model share weights, after they have been appropriately transformed ()[Insert citation for Robert's paper].

We chose to model the discrete choice model using a multinomial logit, however, a generalized linear model (GLM) using the Poisson family with a log-link function would also have been appropriate to estimate the share weights. The multinomial and Poisson models have an equivalence that can be exploited using a generalized linear model.

If we have a set of Poisson distributed random variables, such as the annual capacity additions, $X_1 \sim P(\lambda_1)$, $X_2 \sim P(\lambda_2)$, ..., $X_k \sim P(\lambda_k)$ 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 $\pi_j = \frac{\lambda_j}{\sum_k \lambda_k}$.[12] However, given the well developed mlogit package in R developed by Croissant 2013, in particular the ease in which it implements nested model structures, we chose to implement the multinomial logit model directly.

The logit parameterization makes the assumption that the density $f(\epsilon_{ni})$ in Equation 2 is distributed iid extreme value $\forall\ i$ (or generalized extreme value in the case of the nested logit model). The logit parameterization assumes that the unobserved portion of the utility function are uncorrelated over the choice set. In later model iterations we implement a probit model, which assumes that the unobserved portion of utility (ϵ_n) is distributed multivariate normal. Unlike the logit and nested logit models the resulting integral of the probit model does not have a closed form and thus requires simulation to be evaluated.

In the case of the multinomial logit the density and cumulative distribution function of the unobserved utility is assumed to be, respectively:

$$f(\epsilon_{nj}) = e^{-\epsilon_{nj}} e^{-e^{-\epsilon_{nj}}} \tag{4}$$
$$F(\epsilon_{nj}) = e^{-e^{-\epsilon_{nj}}}$$

[12]The reader is directed to (Guzman lecture notes) for more details on the equivalent log-linear model.

The probability of choosing technology i, which can be interpreted as the joint probability of not choosing alternative i multiplied by the probability of choosing alternative i is then

$$Pr_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-(\epsilon_{nj} + V_{ni} - V_{nj})}} \right) e^{\epsilon_{ni}} e^{-e^{\epsilon_{ni}}} \, d\epsilon_{ni} \tag{5}$$

The integral simplifies to the following:[13]

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \tag{6}$$

As was mentioned earlier, the multinomial logit model imposes the restriction that the error distribution is independent and identical over alternatives (i.e., unobserved portion of the utility is uncorrelated over the choice set and have the same variance) (Wen and Koppelman 2001). The implications of the IIA assumption are that the ratio of two alternatives ($\frac{P_{coal}}{P_n g} = \frac{e^{V_{coal}}}{e^{V_{ng}}}$) is only a function of characteristics of each alternative and not related to other alternatives.

However, given our knowledge of the electricity sector, we expect to see correlation among technologies that use the same primary fuel. To account for the violation of the IIA assumption we could respecify the model to account for the correlation structure; explore other alternatives that do not have this restrictive assumption (i.e., GEV, probit, and mixed logit); or continue with the logit specification as an approximation.

We chose to use the generalized extreme value model (GEV), also known as the nested logit model, which assumes that the error terms are correlated within each nest but independent across nests (see Equation 7). The generalized extreme value model (GEV) allows for correlation in the unobserved factors over alternatives.[14]

$$exp\Big(- \sum_{m=1}^{M} \Big(\sum_{j \in B_m} e^{\frac{-\epsilon_j}{\lambda_m}} \Big)^{\lambda_m} \Big) \tag{7}$$

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

$$P_j = P(j|l)P(l) = \frac{e^{Z_j/\lambda_l}}{\sum_{k \in B_l} e^{Z_k/\lambda_l}} \cdot \frac{e^{W_l + \lambda_l I_l}}{\sum_{m=l}^{M} e^{W_m + \lambda_m I_m}} \tag{8}$$

The link between the branch (i.e., second level) and the twigs (i.e., lower level) is generally referred to as the inclusive utility, which is the expected utility of that nest and is calculated as

$$I_l = ln\Big(\sum_{k \in B_l} e^{Z_k/\lambda_l} \Big) \tag{9}$$

The model is estimated sequentially, with the twigs estimated first and then the branches.

[13]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 $J = 2$.

[14]The GEV model reduces to the multinomial logit if $\lambda_m = 1$.

5. Results

We first test the IIA assumption of the multinomial logit model using the Hausman consistency test as implemented through the 'hmftest' in the mlogit package on the non-nested model structure (Croissant 2013). The Hausman test checks that the probability ratio between each pair of choice alternatives is only determined by the characteristics of the alternative, as opposed to factors of other alternatives. The test compares the full model with a model estimated across a subset of the choice alternatives. The results of our analysis suggest that we reject the null hypothesis of IIA, which means that we should estimate the model with a structure that accounts for the correlated error terms.

To address the violation of the IIA assumption we estimate a nested multinomial model using the nesting structure of the GCAM electricity sector. The nested logit model assumes that the error terms are generalized extreme value (GEV) distributed, which is an extension of the conditional logit model that assumes error terms are extreme value type I, by allowing for correlation among alternatives located in the same nest. The joint distribution of each nest now includes a new term τ_m that represents the correlation of the error terms among alternatives located in the nest (often called the dissimilarity or inclusive value paramter). The choice then among alternatives within each nest remains a conditional logit model. The utility from each alternative is scaled by the inverse of the dissimilarity parameter for the nest.

The nests in GCAM are based on primary fuel use with natural gas, coal, biomass, nuclear, refined liquids, solar, wind, and geothermal, as the available categories. At this time we don't have biomass fuel costs, which resulted in the biomass capacity expansion from being precluded from the model. Further, a nuclear plant has not been built in the United States since the 1970's and since it is an unselected alternative over the period of analysis the model will never select a nuclear plant. The structure of the model then has the following nests: natural gas, coal, oil, geothermal, solar, and wind. The technologies within each nest include coal - coal; natural gas - conventional combustion turbine and combined cycle; oil - conventional combustion turbine and combined cycle; renewables - geothermal, wind, solar thermal, and photovoltaic.

Further, the GCAM electricity sector uses the total cost per each technology, fuel pairing instead of allowing the sub-components of the total cost to vary independently. For equivalence, we generate a new cost term (one per technology, primary fuel pairing for each year) by summing the adjusted overnight cost, adjusted fuel price, variable operations and maintenance, and fixed operations and maintenance costs.

The coal, wind, and geothermal branches are degenerate in that only one option is available in the nest. This can cause estimation issues in the nested logit framework through the dissimilarity index. The conditional logit model implicitly scales each utility so that the error terms have a variance of $\frac{\pi^2}{6}$. In the nested logit model we assume the IIA assumption violated with alternatives located in the same nest now positively correlated. Within each nest the utility resulting from the selection of each alternative is implicitly scaled by the same variance, however, an issue to be addressed is that as the degree of correlation of the error terms increases, the variance decreases within the nest, which can lead to issues of comparing utility across nests. If we normalize each nest by the nest dissimilarity parameter then we can compare the utilities derived from alternatives located in different nests (otherwise we would be comparing apples and oranges, as the utilities associated with each nest would be scaled by different factors)(Heiss et al. 2002).

Greene showed that we can estimate a model with degenerate nests if we fix the dissimilarity parameter for the nest. There is some discussion about whether to fix the numerator or denominator

() but the important thing is that we identify the choice that we made so that models can be compared. We implement the correction for degenerate nests in the mlogit package in R authored by Croissant 2013 by setting the 'unscaled' parameter equal to True. The model specification results in the following raw output (see Table **??**)

TABLE 1. Model 1 output: Nested logit structure to match GCAM

Variable	Estimate	Std. Error	t-value
cost	-0.0001267926	0.0000010466	-121.152
iv.coal	-2.8231532034	0.0587462264	-48.057
iv.ng	8.4443789459	0.0256093450	329.738
iv.oil	2.0880832676	0.0472627419	44.180
iv.solar	-0.3688732253	0.0119622530	-30.836
iv.geothermal	5.1993393429	0.0946018914	54.960
iv.wind	-3.0364914202	0.0465553286	-65.223

The exponentiated parameter estimates are as follows:

TABLE 2. Model 1 output: Exponentiated parameter estimates

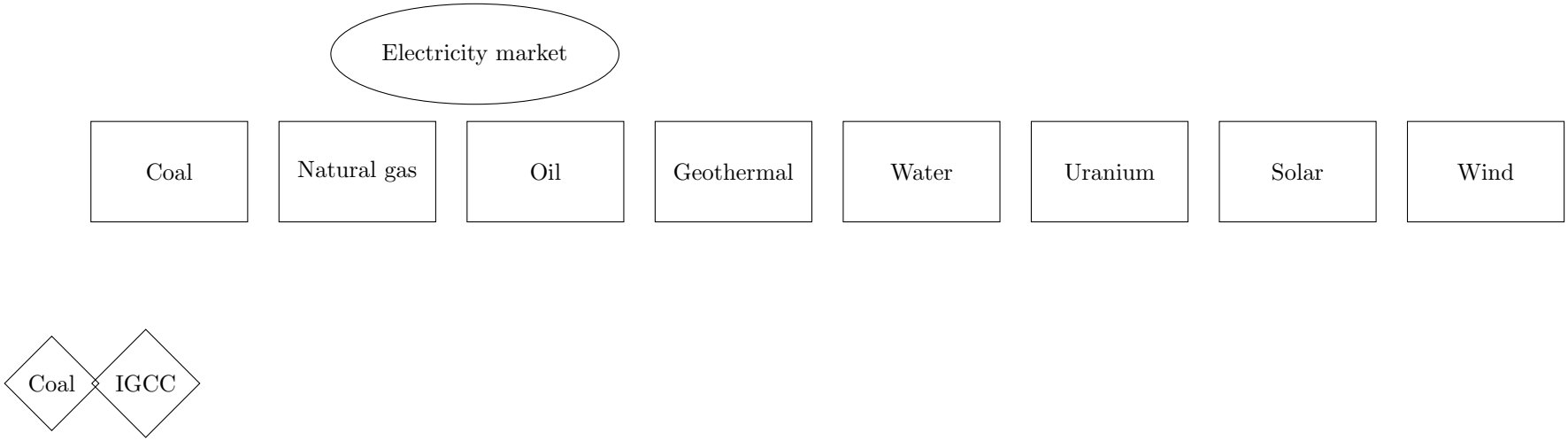
Variable	Estimate
cost	0.99987322
iv.coal	0.05941829
iv.ng	4648.86761788
iv.oil	8.06943339
iv.solar	0.69151307
iv.geothermal	181.15252263
iv.wind	0.04800302

The historic allocation across electricity production technologies can be seen in Table 3.

TABLE 3. Historic allocation of share weights to new capacity additions (1999-2014)

Input fuel	Production technology	Market share
Coal	Coal	6.2%
Natural gas	Combined Cycle	48.7%
Oil	Combined Cycle	0.2%
Natural gas	Combustion Turbine	22.6%
Oil	Combustion Turbine	1.2%
Geothermal	Geothermal	0.2%
Solar	Photovoltaic	2.5%
Solar	Solar Thermal	0.4%
Wind	Wind Turbine	18.0%

6. Discussion


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

References

California Energy Commission (2016). *QFER CEC-1304 Power Plant Owner Reporting Database*. Ed. by M. Nyberg. URL: http://energyalmanac.ca.gov/electricity/web_qfer/.

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.

– (2016a). *Form EIA-860 detailed data*. URL: https://www.eia.gov/electricity/data/eia860/.

– (2016b). *Form EIA-923 detailed data*. URL: https://www.eia.gov/electricity/data/eia923/.

– (2016c). *Glossary*. URL: http://www.eia.gov/tools/glossary/.

Heiss, Florian et al. (2002). "Structural choice analysis with nested logit models". In: *The Stata Journal* 2.3, pp. 227–252.

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.