DOCTORAL THESIS

2019

THREE ESSAYS ON THE SPANISH ELECTRICAL SECTOR

DIEGO JOSÉ BODAS SAGI

PhD PROGRAMME IN ECONOMICS AND BUSINESS

DIRECTOR: JOSÉ MARÍA LABEAGA AZCONA

Department of Economic Theory and Mathematical Economics
Faculty of Economics and Business Administration
Universidad Nacional de Educación a Distancia (UNED)

THREE ESSAYS ON THE SPANISH ELECTRICAL SECTOR

Diego José Bodas Sagi
PhD in System Engineering and Automation
Graduate in Mathematics

DIRECTOR: DR. JOSÉ MARÍA LABEAGA AZCONA

"I believe it is possible that decades from now we will see an era of peace - but we must work together as global citizens of a shared planet" (Dalai Lama)

Acknowledgements

I have been truly lucky to have a wonderful supervisor like José María Labeaga. I am extremely grateful to him for giving me complete freedom in pursuing my interests while also providing research guidance, for sharing with me his tremendous intellectual enthusiasm, and for the countless lessons about research, economy, and life. Thanks José María for your leadership, integrity and commitment.

This work has been possible thanks to the efforts of the UNED, leading the public service of higher education through the modality of distance education. Thanks to this great institution, many students are able to continue studying, learning and keeping growing as a person.

I owe everything to my family. Thanks to my parents for being the perfect example of honesty, self-discipline and sacrificing spirit, fighting with determination against all odds. Thanks to my brother for bringing home joy and jokes, pure oxygen. I cannot forget my true friends and their continuous support. Finally, thanks to my wife and children, the reason of my life, thanks for your love and patience. I am very lucky to have you all.

Abstract

To understand the energy market is paramount to carry out the right policies. The growing demand for affordable, reliable, domestically sourced, and low-carbon electricity is a matter of concern and it is driven by several causes including public policy priorities. In this thesis, the Iberian electricity market from 2011 to 2014 is analyzed to evaluate the price elasticity vis-á-vis changes in certain variables. During this period price elasticity has been estimated. Elasticities are a key factor in understanding how public policies or economic conditions impact the market structure. The work also aims to predict energy demand in Spain for the year 2020 and analyzes whether Spain will be able to meet the European Union's greenhouse gas emission reduction commitment. To this purpose, climatic data and some variables to measure the economic activity in Spain are used. In this country, a series of regulatory reforms since 2015 reduce revenues to existing renewable power generators and they end up the previous system of support to new renewable generation. This policy change has altered the composition of the energy market affecting investment decisions. The public opinion about the national energy policy may be a useful component for investment decisions. For this reason, this sentiment of climate of opinion about energy policy of the Spanish Government is analyzed using the Global Database of Events, Language, and Tone (GDELT).

Resumen

Comprender de forma plena el funcionamiento y dependencias del mercado energético es esencial de cara a poder diseñar las políticas adecuadas que maximicen el interés público. Uno de los objetivos principales de las políticas energéticas actuales tiene que ver con mantener los niveles de desarrollo económico al mismo tiempo que se minimiza la cantidad de gases de efecto invernadero emitidos a la atmósfera y se avanza en el cumplimiento de las nuevas directrices europeas de emisión de gases contaminantes. En este contexto y en este trabajo se analiza el mercado eléctrico español de 2011 a 2014 para evaluar la elasticidad precio frente a los cambios en ciertas variables. Las elasticidades son un factor clave para entender cómo las políticas públicas o las condiciones económicas afectan la estructura del mercado. También se pretende predecir la demanda de energía en España para el año 2020 y analizar si España podrá cumplir el compromiso de reducción de emisiones de gases de efecto invernadero adquirido con la Unión Europea. Para ello, se utilizan datos climáticos y algunas variables para medir la actividad económica del país. Por último, se analiza la opinión pública sobre las reformas regulatorias energéticas introducidas por el Gobierno español en el año 2015, a través de la Base de Datos Global de Eventos, Lenguaje y Tonalidad (GDELT).

Table of contents

List of figures **xiii**

List of tables **xv**

1 Sistema energético español **1**

1.1 Introducción 1

1.1.1 Sistema energético español, contexto 3

1.2 Objetivos del trabajo y metodología 4

1.3 Aportaciones de la tesis doctoral 6

1.3.1 Análisis de la elasticidad de precios y demanda energética en el mercado MIBEL empleando diferentes métodos econométricos . . 6

1.3.2 Estimación de la demanda de energía eléctrica en España y cumplimiento de los acuerdos de emisión de gases de efecto invernadero 7

1.3.3 Usando GDELT para evaluar la confianza de la opinión pública en la política energética del gobierno español 9

1.4 Conclusiones y resumen del capítulo 10

2 Sensitivity of Electricity Price Elasticities **13**

2.1 Introduction 13

2.2 Preliminary description of the data 14

2.3 A set up 18

2.3.1 Model validation 19

2.3.2 Methods 20

2.4 Results and testing 26

2.4.1 Economic results: comparison of elasticities 27

2.4.2 Checking the model from a statistical point of view 31

2.4.3 An additional experiment to robustness 34

2.5 Conclusions and policy implications 36

3 Predicting Energy Demand in Spain **39**
3.1 Introduction 39
3.2 The data 40
3.3 Methodology and results 44
3.3.1 Estimation methods 44
3.3.2 Empirical results 47
3.3.3 Adjusting a rational demand model 53
3.4 Simulating energy demand for 2020 56
3.5 Conclusions 57

4 Using GDELT to Evaluate Confidence **59**
4.1 Introduction 59
4.2 The GDELT Project 61
4.3 Methodology 61
4.3.1 Getting the text of the documents 63
4.4 Results 65
4.4.1 Results using data from GDELT GKG 65
4.4.2 Correlation analysis: Prices and Demand 73
4.4.3 CTM results using CTM algorithm 74
4.4.4 Speedup and Efficiency analysis (getting the text) 75
4.5 Conclusions and future work 77

5 Conclusiones **79**
5.1 Conclusiones y trabajo futuro 79

References **81**

Appendix A Publications **87**

List of figures

2.1 Log price vs log market demand (MIBEL market) 16
2.2 Production by year and technology 16
2.3 Correlations results by energy source 18
2.4 Testing procedure (year 2014) 20
2.5 Variable importance by Random Forest 31

3.1 HDD and CDD evolution over the period 2007 - 2015 42
3.2 IPI evolution for the period 2007 - 2015 42
3.3 GPD and unemployment evolution (2007 - 2015) 43
3.4 GDP adjusted for unemployment (2007-2015) 43
3.5 Energy demand in Spain 43
3.6 Results on test set using linear regression 47
3.7 Results on training set using linear regression 48
3.8 Results on test set using SVR 48
3.9 Results on training set using SVR 49
3.10 Deep Learning Neural Networks applied to training set 50
3.11 Results on test set using Deep Learning Neural Networks 50
3.12 Tuning ε value for SVR and base temperature 16.5 52

4.1 All mentions for "FUELPRICES" theme in Spain 66
4.2 Mentions for "FUELPRICES" theme in Spain filtering words (government mentions) 66
4.3 Histogram - All mentions for "FUELPRICES" theme in Spain 67
4.4 Histogram - Mentions for "FUELPRICES" theme in Spain filtering words (government mentions) 67
4.5 Q-Q Plot - All mentions for "FUELPRICES" theme in Spain 68
4.6 Q-Q Plot - Mentions for "FUELPRICES" theme in Spain filtering words . . 69
4.7 All mentions for "ENV_SOLAR" theme in Spain 70

4.8 Mentions for "ENV_SOLAR" theme in Spain filtering words (government mentions) 70
4.9 Histogram – All mentions for "ENV_SOLAR" theme in Spain 71
4.10 Histogram – Mentions for "ENV_SOLAR" theme in Spain filtering words (government mentions) 71
4.11 Q-Q Plot – All mentions for "ENV_SOLAR" theme in Spain 72
4.12 Q-Q Plot – Mentions for "ENV_SOLAR" theme in Spain filtering words . . 72
4.13 Correlations results 73
4.14 Using "topicmodels" R package. "FUELPRICES" theme, text written in Spanish 74
4.15 Using "topicmodels" R package. "ENV_SOLAR" theme, text written in Spanish 75
4.16 Speedup comparison 76
4.17 Efficiency analysis 76

List of tables

2.1 Variable description 15
2.2 Correlation results 17
2.3 Gene representation (example) 25
2.4 EA parameters 25
2.5 Regression results 28
2.6 Standard deviations of variables for the period 2011 – 2014 30
2.7 Model fit evaluation (predicting MIBEL demand taking logs) 33
2.8 Model fit evaluation for 'fake' data 35

3.1 Deep Learning Neural Network experiments 49
3.2 MSE of different methods 51
3.3 R^2 of different methods 51
3.4 MSE using nominal GDP and GDP deflator 53
3.5 R^2 using nominal GDP and GDP deflator 53
3.6 R^2 using nominal GDP and GDP deflator - taking logs in energy demand . . 54
3.7 MSE adding information about energy prices and CPI 55
3.8 R^2 adding information about energy prices and CPI 55
3.9 MSE adding information about nominal GDP and population growth rate . 55
3.10 R^2 adding information about nominal GDP and population growth rate . . . 56

Chapter 1

Sistema energético español

1.1 Introducción

El matemático Anfred James Lotka en su obra Elements of Physical Biology ([2]) concluyó que los organismos que prosperan y sobreviven son aquellos que usan la energía de forma más eficiente y efectiva que sus competidores. Posteriormente, Howard Odum reformula las hipótesis de Lotka estableciendo que en un proceso auto-organizado los sistemas desarrollan aquellas partes, procesos y relaciones que capturan la mayor cantidad posible de energía y lo usan de la forma más eficiente posible sin reducir la potencia generada ([45]).

Ambos autores nos llevan a pensar en varios aspectos relacionados con la producción y consumo energético. En primer lugar, hay una relación directa entre la supervivencia de una especie y su capacidad para aprovechar la energía que captura del entorno. En segundo lugar, el uso eficiente de la energía es un imperativo para la propia supervivencia.

Desde el punto de vista de la Física, la eficiencia energética indica el porcentaje de la energía total de entrada a un sistema que es consumida en forma de trabajo útil y no se desperdicia. Desde un punto de vista más económico, se enmarca dentro de un programa de eficiencia energética el conjunto de estrategias y acciones destinadas a reducir el consumo de energía de un sistema sin que se vea afectada la calidad o cantidad del servicio suministrado ([65]). Las cuestiones medioambientales son hoy en día un punto de interés esencial en las políticas de eficiencia energética como más adelante se mencionará.

Toda sociedad, como un sistema más, necesita energía para su mantenimiento y evolución. La energía actúa como factor limitador del desarrollo. La política energética de los gobiernos e instituciones tiene un impacto destacado en el desarrollo en general de una sociedad y en el desarrollo económico en particular. Corresponde a la política energética de un país determinar la estrategia de producción energética, distribución y consumo con el fin de asegurar el suministro eficiente de energía a ciudadanos, empresas e instituciones. Las

herramientas de las que dispone incluye desde las leyes hasta los incentivos a la inversión, pasando por la firma de tratados internacionales. Dado que el precio juega un papel importante en la demanda del bien, la política impositiva tiene un papel destacado a la hora de actuar sobre la demanda ([29]).

En un contexto como el actual de sobreexplotación de recursos naturales, las cuestiones medioambientales juegan un papel fundamental a la hora de diseñar la política energética de un gobierno. En este sentido, la Comisión Europea patrocina un programa de investigación dotado de más de 2 mil millones de euros y dedicado al fomento de la investigación en energía, medio ambiente y desarrollo sostenible[1]. Cuestión no trivial si tenemos en cuenta que, según Naciones Unidas, se espera que la población mundial alcance una cifra superior a 9 billones en el año 2050 y de más de 10 billones en el año 2100.

Los acuerdos de la EU obligan a que para el año 2020 entre el 20% y 30% de la energía total consumida provenga de fuentes limpias y renovables. El objetivo es recortar la emisión de gases de efecto invernadero un 20% respecto a los niveles de 1990. Así se estipula en los artículos 17, 18 y 19 de la Directiva EU 2009/28/CE del Parlamento Europeo y Consejo del 23 de abril de 2009. La norma se traslada a España en el Real Decreto 1597/2011 del 4 de noviembre de 2011. En diciembre de 2015, dichos acuerdos son ratificados en la conferencia contra el cambio climático de Naciones Unidas. Posteriormente, es el Parlamento Europeo en sesión del 4 de octubre de 2016 quien se reafirma en el compromiso.

El centro de investigación en materia energética Global Energy Assessment (GEA) enmarcado dentro del International Institute for Applied Systems Analysis (IIASA), surge con la intención de analizar la política energética a nivel mundial y proporcionar ideas a gobiernos, empresas y organizaciones de todo tipo para el desarrollo de una política energética sostenible y que, al mismo tiempo, sea capaz de lidiar con los retos y oportunidades que el siglo XXI abre. Ya en el año 2012 el centro lanzó una serie de recomendaciones destinadas a los gobiernos de todos los países. Se proponía una amplio campo de acción que abarcaba desde trabajar para asegurar el suministro energético, balanceando correctamente las importaciones con las exportaciones, hasta impulsar un desarrollo energético sostenible con un menor impacto medioambiental. Según GEA, llevar a cabo estas propuestas es factible porque la energía renovable es abundante y los costes de generación son escalables. Además, la evolución de la tecnología puede facilitar el proceso. Este paradigma energético podría ser una fuente de generación de empleo y de nacimiento de nuevas oportunidades económicas. Bien es cierto que, para que se produzca, se precisa de unas directrices políticas claras que desemboquen en un sistema regulatorio adecuado y que impulsen, a través de

[1] http://cordis.europa.eu/programme/rcn/643_en.html

planes educativos, un cambio de mentalidad que refuerce la conciencia medioambiental y el consumo energético responsable.

Dado que las cuestiones políticas tienen una gran influencia sobre la evolución del mercado energético, es preciso disponer de herramientas eficaces y eficientes que permitan medir y analizar el impacto de las políticas públicas. Esta evaluación posibilita la mejora continua del proceso de toma de decisiones. La presente tesis se enmarca en este contexto, proponiendo nuevos mecanismos y procedimientos para estudiar la evolución de precios, demanda y emisión de gases contaminantes.

1.1.1 Sistema energético español, contexto

El sistema energético español debe dar servicio a una superficie de más de 500.000 m^2, 46 millones de personas y un clima diverso. Predomina el clima mediterráneo con temperaturas cálidas pero también posee zonas con clima continental, oceánico, clima de montaña y subtropical. El PIB por habitante se sitúa en torno a los 23.000 euros en el año 2016.

A la hora de entender el contexto energético español debemos tener en cuenta que España, como miembro de la Unión Europea, está obligada al cumplimiento de las directrices europeas. El tercer paquete energético de la EU que entra en vigor en septiembre de 2009, está compuesto por tres reglamentos (Reglamento (CE) Nº 713/2009, Reglamento (CE) Nº 714/2009 y Reglamento (CE) Nº 715/2009) y dos directivas (Directiva 2009/72/CE y Directiva 2009/73/CE). En el caso de los reglamentos, la base jurídica es el artículo 95 del anterior Tratado de la Unión Europea (TUE), mientras que en el caso de las Directivas se trata de los artículos 47(2), 55 y 95 TUE. El objetivo de la propuesta es la separación de las actividades de producción y distribución como medio para fomentar la competencia y la inversión en infraestructuras. Dentro de un mercado, la competencia favorece la innovación y beneficia al consumidor ya que le posibilita un amplio rango de elección. Sin embargo, la Comisión constató en el año 2016, que la actual legislación española impide a las empresas distintas de los gestores históricos nacionales de las redes de gas y electricidad construir y explotar interconexiones con otros Estados miembros[2].

El déficit de tarifa es una de las peculiaridades propias del mercado español. Se denomina déficit de tarifa a la diferencia entre los ingresos obtenidos por los precios regulados que pagan los consumidores y los costes reales de dicho suministro. El cómputo se inicia en el año 2001. Parte de la diferencia puede deberse a posibles errores de estimación y objetivos políticos/económicos de los sucesivos gobiernos. Entre los años 2005 y 2013, los costes del sistema eléctrico se incrementaron en tasas superiores al 200%. Los beneficios no alcanzaron

[2]http://europa.eu/rapid/press-release_MEMO-16-3125_es.htm

esa cuota, llegando a incrementarse un 100%. Las subvenciones al sistema de energía renovable constituyeron uno de los mayores conceptos de gasto. En el año 2012, la deuda acumulada del sistema alcanzó los 20 millones de euros, momento en el que las subvenciones a las energías renovable fue abolida. Al mismo tiempo, se redujeron las cantidades que el ejecutivo abona en concepto de tareas relacionadas con la transmisión y distribución de la energía. A pesar de todo, a finales del mismo año 2012 el déficit era de 26 millones de euros. En julio de 2013, el gobierno inicia un amplio paquete de reformas buscando una reducción de costes. El precio que los consumidores finales españoles abonan por la electricidad que consumen está entre los más altos de Europa. Los distintos planes nacionales de eficiencia energética tienen como objetivo la reforma del mercado eléctrico de cara a asegurar la sostenibilidad y proteger a los consumidores.

La Agencia Internacional de la Energía (IEA) en el año 2015 recomienda las siguientes acciones al gobierno español:

- Introducir un sistema coherente que permita balancear correctamente las distintas fuentes de energía eléctrica buscando maximizar la eficiencia energética y minimizar la emisión de gases de efecto invernadero. Hace especial hincapié en controlar estos aspectos una vez se origine un nuevo crecimiento continuado de la actividad económica.
- Reducir el consumo y venta de combustibles fósiles al menos en un 1.5% para el año 2020.
- Obligar a las empresas a implementar proyectos de eficiencia energética. Parte de estos proyectos pueden ser financiados en virtud de la ley 18/2014 del 15 de octubre.

1.2 Objetivos del trabajo y metodología

En este trabajo se analizan diferentes aspectos relacionados con la producción y demanda de energía eléctrica. Se restringe el área de estudio a la Península Ibérica y de forma más concreta a España, cuyas actuaciones energéticas deben enmarcarse dentro de la directrices en materia energética de la Unión Europea. Los objetivos y actuaciones políticas así como las innovaciones tecnológicas impactan sobre el mercado eléctrico generando cambios en la estructura del mercado y afectando a productores, distribuidores y consumidores. Tres son los objetivos fundamentales que se acometen en la presente tesis:

1. Analizar la relación entre precio y demanda de energía eléctrica en el mercado energético de la península Ibérica. Se pretende concluir cuáles son las variables que más influyen en los incrementos o decrementos del precio.

2. Proponer mecanismos que permitan evaluar si España será capaz de cumplir con las obligaciones incurridas con la EU en lo que se refiere a la reducción de gases de efecto invernadero.

3. Analizar desde una perspectiva técnica la opinión que la comunidad pública ha manifestado sobre ciertas actuaciones del gobierno español en materia energética, contribuyendo a la difusión de técnicas que pueden ayudar a una mejor evaluación de la opinión que el mercado tiene sobre las acciones del ejecutivo.

De forma más detallada, se comentan a continuación los tres objetivos planteados. En primer lugar, se estudia la relación entre diversas variables relacionadas con el mercado energético, dos de las variables más importantes son la demanda de energía eléctrica y precio. Se emplean datos correspondientes al periodo comprendido entre los años 2011 y 2014. En segundo lugar, se plantea el uso de variables económica y climatológicas para intentar constatar si España será capaz de cumplir con los compromisos adquiridos con la EU en materia de emisión de gases de efectos invernaderos. Debido a que, basándose en datos históricos, la probabilidad de un incremento en la demanda energética aumenta en periodos de auge económico, se considera un escenario de crecimiento económico continuo hasta el año 2020 manteniendo las condiciones climatológicas. Por último y de cara a proponer alternativas para recoger y analizar de forma algorítmica y automática la opinión que la opinión pública tiene sobre ciertas medidas, se analiza información disponible digitalmente para intentar evaluar el sentimiento del mercado sobre una de las medidas más comentadas del gobierno español en materia energética, nos referimos a la regulación comúnmente conocida como “impuesto al sol”.

Los tres objetivos mencionados anteriormente se han abordado de forma iterativa e independiente. Para todos los casos y en base al objetivo concreto planteado, se ha procedido a la búsqueda de fuentes de información con la intención de recopilar datos susceptibles de aportar valor al estudio. Los datos han debido superar un proceso de limpieza, transformación y normalización. A continuación, diversas técnicas analíticas han sido propuestas, contrastando los resultados obtenidos por cada una de ellas. Se ha dado preferencia a modelos interpretables, aunque también se deja la puerta abierta al uso de modelos cuya composición final no es directamente interpretable como sucede en el caso de las redes neuronales de varias capas. Con el fin de evitar el sobreajuste y obtener modelos con unos ratios adecuados de fiabilidad predictiva, se ha sometido el modelo a procesos de entrenamiento y validación empleando conjunto de datos distintos. Por último y para cada caso, se exponen las conclusiones finales y posibilidades de ampliación.

El resto del trabajo se divide como sigue; en los apartados siguientes del presente capítulo resumimos las aportaciones de la tesis doctoral y principales conclusiones. El capítulo 2 analiza la relación entre precio de la energía y demanda de la misma en la península Ibérica para el periodo 2011-2014. El tercer capítulo propone un procedimiento para intentar determinar las emisiones de dióxido de carbono a la atmósfera en el año 2020 dentro de un contexto determinado. Por último, el capítulo final propone una metodología que permite recopilar la opinión pública a nivel global asociado a determinados entes como instituciones o gobiernos nacionales.

1.3 Aportaciones de la tesis doctoral

En esta tesis doctoral se enfocan tres ensayos relacionados cuyas características se exponen a continuación.

1.3.1 Análisis de la elasticidad de precios y demanda energética en el mercado MIBEL empleando diferentes métodos econométricos

En este trabajo se pretende avanzar el conocimiento del mercado del mercado energético de la península Ibérica analizando la relación entre precio y demanda de energía eléctrica. Se pretende concluir cuáles son las variables que más influyen en los incrementos o decrementos del precio. Los datos abarcan el periodo comprendido entre los años 2011 y 2014, estando disponibles en intervalos horarios y habiendo sido agregados diariamente para cumplir con el propósito del trabajo. No solo se analiza la relación entre precio y demanda sino que también se ponen en juego otro tipo de variables analizando el impacto que cada una de ellas tiene sobre la demanda total energética de energía eléctrica. Empleando y contrastando entre sí varias técnicas de regresión, se pretende proporcionar datos que sean de utilidad a la hora de fijar políticas económicas que fomenten o desincentiven el empleo de ciertas tecnologías de producción según el objetivo perseguido.

El análisis de los resultados incide no sólo en la comparación del poder predictivo de las distintas técnicas ante el modelo planteado, sino que también se analizan las diferentes elasticidades proporcionadas por los distintos métodos evaluados. Estas elasticidades son un elemento esencial de cara a comprender cómo las políticas públicas o las condiciones económicas impactan en la estructura del mercado. En el caso de uso concreto seleccionado, se ha demostrado que los consumidores (interpretados a través de la demanda) no reaccionan de forma significativa a las variaciones de precios recogidas en el histórico de datos. El

capítulo incluye una sección dedicada a corroborar las fortaleza de los resultados y a justificar la necesidad de emplear varias métricas de validación.

Varias han sido las técnicas empleadas para la construcción de los modelos matemáticos. En primer lugar, se ha empleado un modelo clásico de regresión lineal ([34]) con el objetivo de minimizar el error de mínimos cuadrados ordinarios ([56]). En segundo lugar, se han evaluado modelos de regresión Ridge ([32]; [34]) y Lassso (Least Absolute Shrinkage and Selection Operator) ([59]). Estas técnicas se han complementado con otras procedentes del ámbito del aprendizaje automático. En concreto, las técnicas dentro de este campo que también se han evaluado son: Support Vector Machines (SVM) ([16]; [53]), en su vertiente Support Vector Regression (SVR) ([12]), Algoritmos evolutivos ([66]; [25]), y Random Forest ([8]). Esta última técnica, se ha empleado para baremar la importancia de las distintas variables empleadas.

El estudio concluye demostrando que el modelo OLS obtiene mejores resultados que otras opciones probadas, como SVM, Lasso o Ridge. Estos dos últimos modelos no contribuyen a diferenciar de forma significativa la contribución de unas variables frente a otras. El modelo basado en SVM (y con la implementación seleccionada), ha encontrado dificultades en ciertos casos debido a un valor bajo en la varianza de la demanda energética. En cuanto al modelo basado en algoritmos evolutivos, también encuentra dificultades para destacar a unas variables frente a otras y salir de óptimos locales.

En el aspecto puramente económico, se encuentra una muy leve reacción de los consumidores a los movimientos en los precios, de forma que baja el consumo ante subidas de precios. Sin embargo, no se han encontrado evidencias estadística significativa sobre esta cuestión. En cuanto a la importancia de las distintas variables, las dos variables más destacadas para predecir la demanda son el precio y la variable que cataloga el día en cuestión en laborable o festivo.

1.3.2 Estimación de la demanda de energía eléctrica en España y cumplimiento de los acuerdos de emisión de gases de efecto invernadero

En el tercer capítulo y en relación al segundo objetivo planteado en esta tesis, se sugiere el uso de variables económicas y climatológicas para, empleando el mínimo de información posible, lanzar una predicción para la demanda de energía eléctrica en España en el año 2020 e intentar así constatar si este país será capaz de cumplir con los compromisos adquiridos con la EU en materia de emisión de gases de efectos invernaderos. Como se mencionaba en la introducción, estos acuerdos implican que en el año 2020 entre el 20% y el 30% de la

energía consumida, debe provenir de energías renovables. En este caso, el conjunto de datos recopilados abarca desde el año 2007 al año 2015.

Las variables elegidas son el Índice de Producción Industrial (IPI) publicado por el Instituto Nacional de Estadística (INE) además de variables climatológicas. El IPI mide la evolución mensual de la actividad productiva de las industrias extractivas, manufactureras y de producción y distribución de energía eléctrica, agua y gas. A partir del año 2010, también considera las industrias relacionadas con la captación, depuración y distribución de agua. La elección de esta variable se debe a que, tradicionalmente, el índice IPI está correlacionado con el gasto anual medio de las familias españolas ([40]). Por esta razón, se considera que un crecimiento destacado en la actividad económica afectará positivamente a la evolución del índice. Adicionalmente y para evaluar la robustez del modelo, se ha trabajado también con el Producto Interior Bruto (PIB). En cuanto a las variables climatológicas y a través de varias fuentes como la Agencia Estatal de Meteorología de España, Red Eléctrica Española (REE), Ministerio de Agricultura Español y el National Climatic Data Center de los Estados Unidos, se recopilan datos diarios de temperatura y precipitaciones de varias estaciones meteorológicas. Los datos son agregados para proporcionar en el conjunto de la península Ibérica y en el intervalo seleccionado, la temperatura máxima diaria, temperatura mínima diaria, temperatura media diaria y precipitaciones totales diaria. Las islas Baleares y Canarias no han sido tenidas en cuenta en el estudio. Para incorporar las temperaturas a los distintos modelos planteados, se ha empleado la convención Heating y Cooling Degree Days ([49]).

El modelo base planteado propone estimar la cantidad diaria de energía eléctrica demandada en base un modelo lineal que considera datos de temperaturas, precipitaciones y evolución del IPI. Posteriormente, se elimina la restricción del modelo lineal y, manteniendo las mismas variables, se inserta un modelo no lineal proporcionado por una red neuronal.

Las conclusiones reflejan que, en un escenario de crecimiento económico continuo hasta el año 2020 y mantenimiento de las condiciones climatológicas, España será capaz de cumplir con sus obligaciones. Esto coincide con los mensajes lanzados por REE, organismo que, por ejemplo, en julio del año 2015 ya destacaba que, aproximadamente, el 30% de la energía total producida procedía de energías renovables. Sin embargo, el trabajo no permite diluir todas las dudas respecto a la cantidad de CO_2 que podría liberarse a la atmósfera ya a la dificultad para encontrar datos abiertos consistentes, oficiales y fiables en cuanto a la cantidad de CO_2 emitido por kilovatio hora consumido, se une a la dificultad para evaluar, a futuro, la mejora en cuanto a eficiencia y reducción de gases contaminantes de tecnologías de generación eléctrica.

1.3.3 Usando GDELT para evaluar la confianza de la opinión pública en la política energética del gobierno español

Las políticas públicas juegan un papel fundamental regulando la relación entre empresas, inversores y consumidores. Esta influencia ha ido aumentando en los últimos años especialmente en las inversiones a largo plazo debido a la crisis financiera global desatada a partir del año 2008 y a las interdependencias de la actividad de las compañías con el sector medioambiental entre otros ([57]). Los inversores también debe considerar los últimos avances tecnológicos en materia energética siendo, precisamente, la regulación de estas tecnologías una cuestión no trivial objeto de variada controversia. Como ejemplo destacado y foco de estudio, se cita el Real Decreto (RD) de octubre de 2015 emitido por el ejecutivo español. En este RD se oficializa el comúnmente conocido "impuesto al sol", mediante el cual el consumidor que opte por el autoconsumo tendrá que abonar unas cantidades en concepto de conexión o uso de la red de distribución eléctrica. Fue esta una medida fuertemente criticada por asociaciones de consumidores y medioambientales.

Con motivo de esta discutida norma, se pretende analizar mediante un procedimiento riguroso la opinión que la opinión pública en general posee sobre la política energética del gobierno español y, en concreto, sobre el mencionado "impuesto al sol" ejemplo de actuación en, especialmente, energía solar. Para ello, se emplea la base de datos GDELT ([27]) como fuente de información masiva. Toda esta información reside en un repositorio de datos en la nube, con altas prestaciones de cómputo. Aquí se acumulan todo tipo de registros, noticias, entradas en blog, notas de prensa, publicados en internet (exceptuando aquellos contenidos asociados a temática deportiva y ocio). La información se encuentra clasificada en diferentes categorías e incorpora información sobre la tonalidad del contenido, tonalidad que puede moverse desde un rango muy negativo a muy positivo.

De todos los contenidos disponibles en la base de datos GDELT, se han seleccionado aquellos que hacían referencia a cuestiones relacionadas con energía solar o precios del combustible entre las fechas 2013 a 2016, puesto que en fechas previas al año 2013 la información que se necesita no está completa. Además, se ha exigido que el contenido hiciese mención directa a España. Sin embargo, estos filtros demostraron ser insuficientes puesto que mostraban textos irrelevantes para el objetivo de estudio. Por ejemplo, noticias sobre una nueva publicación científica del instituto de astrofísica de Canarias en relación con la energía solar, evidentemente ubicada en España y, además, con una tonalidad muy positiva, se encontraba entre la información a considerar a pesar de no mencionar al gobierno ni cuestiones políticas en ningún momento (pero al fin y al cabo de trataba de un noticia localizada en España y relacionada con temas de energía solar). Es por ello por lo que se ha tenido que profundizar en el análisis de texto añadiendo filtros adicionales a los anteriores.

En concreto, estas capas adicionales consistieron seleccionar únicamente textos escritos en castellano o inglés y conteniendo palabras relacionadas con la búsqueda que interesaba, palabras tales como gobierno, ejecutivo, ministro, Soria (apellido del ministro que aprobó la norma), etc. En estas condiciones, una exploración manual pero exhaustiva verificó que, los textos resultantes sí se referían a cuestiones relacionadas con el sistema energético español, inversión en empresas de energía, actuaciones del gobierno, etc. Dado el volumen del contenido a analizar, fue necesario paralelizar computacionalmente la ejecución de este proceso.

Con los contenidos resultantes se realizó un análisis estadístico básico de la tonalidad de los mensajes comparando entre sí aquellos que incluían términos relacionados con el ejecutivo y frente al total disponible para la categoría seleccionada (contenidos relacionados con la energía solar y que mencionen España). Además, se aplicó un algoritmo de modelado de tópicos ([28]) sobre los textos escritos en castellano. Este algoritmo no supervisado permite agrupar documentos en función de sus términos.

El estudio de los registros refleja que los mismos exponen una opinión negativa sobre la reforma energética del gobierno. Dada las fechas de publicación de los textos analizados, asociamos esta tonalidad a una opinión negativa sobre el llamado "impuesto al sol" instaurado por el gobierno español en octubre de 2015. El estudio también demuestra que existe una correlación muy débil entre el tono de las distintas menciones y el precio diario de la energía. En cuanto a los resultados del algoritmo de modelado de tópicos, se comprueba que la técnica permite clasificar los documentos en distintos grupos. Proponiéndose para trabajos futuros emplear esta información para realizar un estudio más amplio y segmentado en función de los grupos detectados.

El trabajo realizado puede aportar valor a la hora de establecer mecanismos de evaluación de las políticas públicas en materia energética.

1.4 Conclusiones y resumen del capítulo

La presente tesis proporciona información sobre la evolución del mercado eléctrico de la península Ibérica entre los años 2011 y 2015 empleando un amplio conjunto de variables de distinto tipo. Se justifica una débil correlación entre precio y demanda al mismo tiempo que se encuentra relación inversa entre el incrementos de precio y el incremento de la energía eléctrica generada mediante centrales hidroeléctricas, la relación es directa en el caso de caso de electricidad procedente de centrales térmicas.

Además, proporciona una estimación sobre la cantidad de dióxido de carbono susceptible de emitirse a la atmósfera en un contexto de crecimiento económico en el año 2020.

Afirmando que España se encuentra en condiciones de cumplir con los compromisos adquiridos con la Unión Europea en esta fecha y en las condiciones que refleja el estudio.

En un mundo como el actual caracterizado por una generación de datos masivos, el empleo de fuentes de información abiertas como GDELT aporta valor a estudios de opinión sobre distintos temas, pudiendo esto contribuir al análisis de las políticas públicas.

Todo ello lo hace comparando entre sí varias técnicas estadísticas y de aprendizaje automático. Dado el estado del arte, la evolución continua de los algoritmos de aprendizaje automático y en relación con el teorema "no free lunch" ([64]), cada análisis o estudio debe contrastar entre sí los resultados de diferentes modelos de enfoques variados con el fin de encontrar aquél que mejor se adapta a las características del problema.

Chapter 2

Sensitivity of Electricity Price Elasticities to Different Econometric Methods

2.1 Introduction

The development of a country depends heavily on the efficiency of its energy system. As one of the most important components of an energy system, we need to understand the functioning of the electricity markets. Electricity results from the conversion of primary fuels such as fossil fuels, uranium, water, wind or solar into a flow of electrons used to power modern life ([15]). It can be considered as a commodity capable of being bought, sold and traded several times from the initial producer to the end user. As emphasized in [29], all these operations involve commercial arrangements for energy and capacity trading between participants and the system operator. The coordination of such commercial arrangements between players takes place within the electricity markets. These markets involve physical elements (natural resources, infrastructure, institutions, companies...) and financial elements including purchasing/selling of economic products ([15]).

In Spain, OMI-Polo Spanish S.A. is a company regulated by the International Convention of Santiago, in the creation of an Iberian electricity market (MIBEL) between the Kingdom of Spain and the Republic of Portugal, and subject to the electricity sector regulation. The Iberian Market Operator (OMIE) is a company that manages the markets for the whole of the Iberian Peninsula, and its operating model is the same as the one applied by many other European markets.

In this paper, we pursue two main goals. First, we aim to evaluate the response of some variables to changes in prices in the MIBEL. We study the relationship between price and market demand of electricity generated through different energy sources. The present study

provides information on the evolution of Spanish electricity market from 2011 to 2014 and contributes to a better understanding of the market behavior. Estimating price elasticity is a key factor in understanding how public policies can be used to modify individual/corporate behavior, or even forecasting ([17]). Although the focus for estimating the model remains in the period 2011–2014, we use data for 2015 to do additional tests. The second purpose of this paper is to compare several techniques for estimating multi-variable linear regression models. We do not only compare routines and statistics, but we also test the sensitivity of the elasticities obtained with different estimation methods. Although our main interest lies in the sensitivity of the price elasticity, we also compare the performance of each algorithm. As expected, the results show that price increases drive demand reductions. However, its potential is weak. The use of alternative methods to replace more commonly used regression methods does not significantly improve the results. The paper is structured as follows: section 2 describes the dataset used here. Section 3 introduces the implemented methodology and algorithms. Section 4 shows the main results. Finally, the last section concludes and summarizes.

2.2 Preliminary description of the data

The data used in the paper comes from OMIE daily markets (MIBEL), and is available in csv format at ([21]). The original information corresponds to hourly data, however for our purposes it has been aggregated to daily data. We express the variable in logs so, we first take logs in the original data and, then, we obtain the daily average. Our period of analysis spans from 2011 to 2014. Table 2.1 reports the main variables and attributes considered in the econometric specifications.

Table 2.1 Variable description

VARIABLE	***DESCRIPTION***
Date	Format: YYYY-MM-DD
Price	EUR/MWh (taking natural logs, calculated with the mean log price in the day)
Hydro	Daily hydroelectric energy demanded (MWh)
Nuclear	Daily nuclear energy demanded (MWh)
Coal	Daily electricity energy demanded (MWh) using coal as the primary fuel
ComCycle	Daily electricity energy demanded (MWh) by combined-cycle power plants
FuelGas	Daily electricity energy demanded (MWh) using fuel and gas
RegEsp	Daily electricity energy demanded (MWh) by special regime technologies. The application of the Spanish special regime is discretionary for companies that own eligible facilities. Generally, eligible facilities are those with an installed capacity of 50 MW or less that use cogeneration or any renewable energy source as their primary energy
Import	Daily imported electricity energy (MWh)
Export	Daily exported electricity energy (MWh)
LogMarketDemand (MIBEL)	Daily electricity demand (total) on the MIBEL market (MWh) (calculated with the mean log hourly market demand in the day)

In the figure below, we show a crude graphical description of the two key variables we are going to focus on in this study: market demand (in logs) versus log prices. A first feature of this graph is the different volatility of the two variables. According to it, market demand has remained relatively stable over the period 2011-2014. Nevertheless, it appears that log-prices has suffered a slight decrease in the same period, motivated, among other reasons, by the economic crisis.

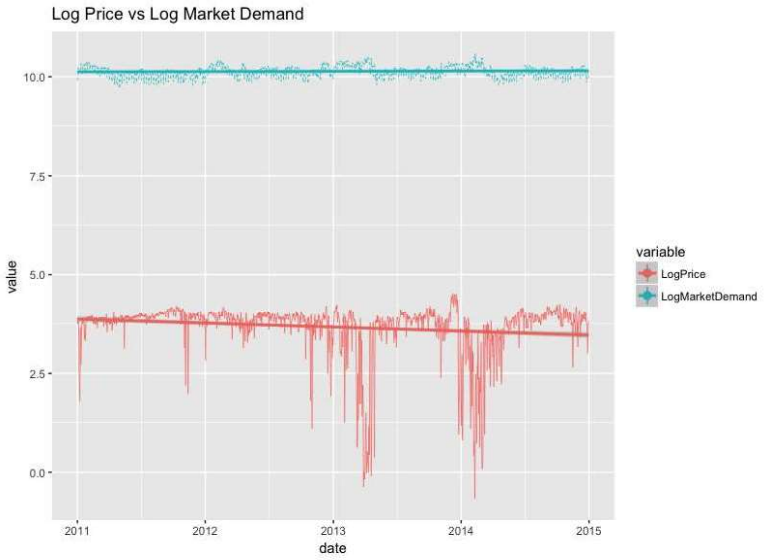

Fig. 2.1 Log price vs log market demand (MIBEL market)

Energy sources grouped by year and technology are shown in the Figure 2.1. The application of the Spanish special regime is discretionary for companies who own eligible facilities. In general, eligible facilities are those with an installed capacity of 50 MW or less which use cogeneration or any renewable energy source as their primary energy.

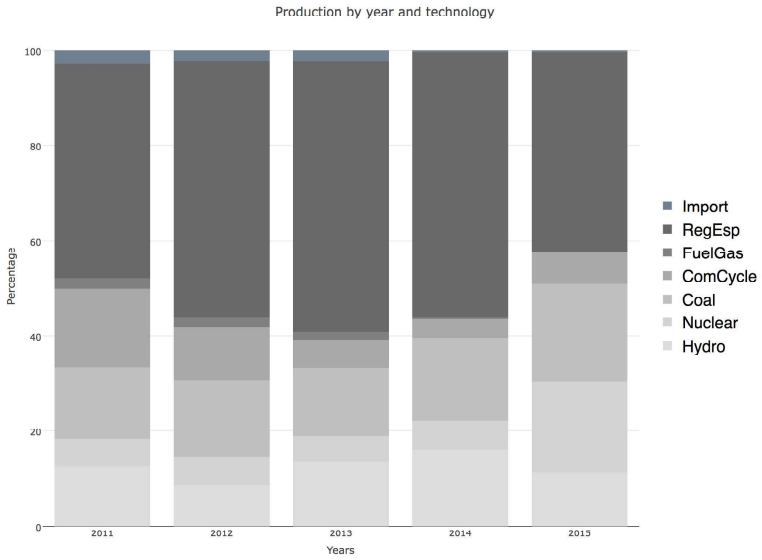

Fig. 2.2 Production by year and technology

Hydroelectricity production is directly correlated with the amount of precipitation in the given year, and this can be seen in the last figure. Year 2014 is the rainiest one, while year 2015 and 2012 are the driest years. Electricity from nuclear sources remain stable in the period 2011-2014, the value increases in the year 2015, maybe, trying to offset the decline of energy produced by special regime technologies. Energy from coal is stable in all the period, whereas energy from fuel and gas suffers a gradual reduction in the period 2011 – 2014, the production increases slightly in the last year.

Correlation table (including significance values) can be found below. Some relevant values have been found, as the relation between special regime and coal or combined-cycled power. Also is high the correlation value between fuel-gas source and imported electricity (Spain does not have important reserves of natural gas).

Table 2.2 Correlation results

LogPrice	***LogPrice***							
LogMarketSupply	-0.27****	***LogMarketSupply***						
Hydro	-0.43****	0.14****	***Hydro***					
Nuclear	-0.57****	0.13****	0.22****	***Nuclear***				
Coal	0.56****	-0.04	-0.43****	-0.36****	***Coal***			
ComCycle	0.39****	0.03	-0.27****	-0.20****	0.32****	***ComCycle***		
FuelGas	0.11****	-0.09**	-0.15****	0.02	-0.14****	0.44****	**FuelGas**	
RegEsp	-0.33****	-0.01	-0.10***	0.05*	-0.67****	-0.66****	-0.11****	***RegEsp***
Import	0.22****	-0.25****	-0.14****	-0.12****	-0.07*	0.34****	0.71****	-0.13****

**** p-value <0.0001
*** p-value <0.001
** p-value <0.01
* p-value <0.05

The same results are now showed in the next figure:

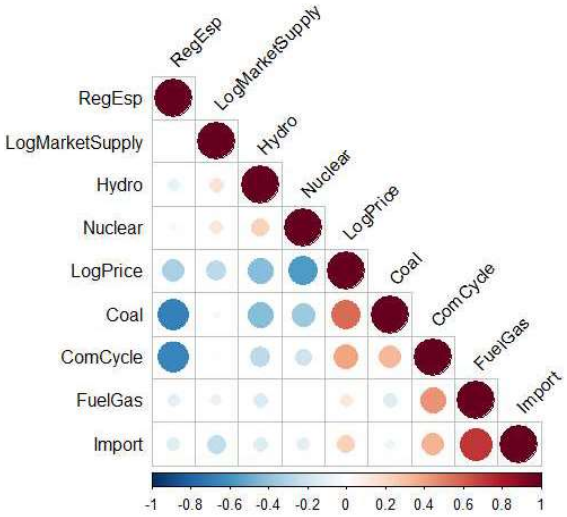

Fig. 2.3 Correlations results by energy source

Although this paper mainly uses data from the years 2011-2014, some experiments have been performed using data from 2015. For this reason and for descriptive purposes, the structure of the different technology sources from this year is included in the last figure. Energy produced from nuclear sources substantially increases in 2015. On the other hand, while electricity offered by the special regime decreases in 2015, imported electricity is close to zero during that period.

2.3 A set up

This section introduces the model we are interested in estimating. It is a basic log-log linear specification for the relationship between demand and prices. We choose this model because of its linear structure and ease for being estimated and compared. Moreover, in a log-log linear demand equation the estimated coefficients are directly the elasticities.

With these conditions, the following linear equation was set up:

$$y_t = \alpha + \beta x_t + \gamma' z_t + u_t \tag{2.1}$$

Where x_t is log price at day t , y_t is log market demand at day t , z_t represents the list of monthly dummies plus a labor day variable, α, β and γ are parameters and u_t is the error.

The original variables are measured on an hourly scale. We produce an aggregate to obtain daily data and calculate the average of any given day. Furthermore, since the system

can have either superavit or deficit, it needs to import or export energy. For this reason, we define two dummies collecting the situation of importing and exporting. Therefore, the specification considers the following additional daily variables:

- Import: Dummy (0/1) variable. If any amount of electricity has been imported that day, the value is 1; otherwise, the value is 0
- Export: Dummy (0/1) variable. If any amount of electricity has been exported that day, the value is 1; otherwise, the value is 0
- We use 12 monthly dummies, for example, for day t the variable Month_i equals to 1 if $\text{month}(t) = i$. For day t , Month_j equals 0 if $\text{month}(t) \neq j$.
- We also include a WeekDay variable such that $\text{WeekDay}(t) = 1$ if $\text{DayOfWeek}(t)$ in $\{Monday, Tuesday, Wednesday, Thursday, Friday\}$; otherwise $\text{WeekDay}(t) = 0$.

2.3.1 Model validation

In order to choose the model that best fits the data (in economic and statistical terms), we are going to perform the following procedure. Given a dataset for one year (i.e., for year 2014) and given a regression technique, we first estimate a linear model using data for that year as the training set. The model for a particular year is tested against the other years. To choose among the different alternatives we compare the models by means of a Minimum Squared Error (MSE) criteria in the test sets. Taking one test set per year, allows us to analyze in a clearer way how the trained model is performing in other years. The following figure can help understand the procedure, taking as an example year 2014.

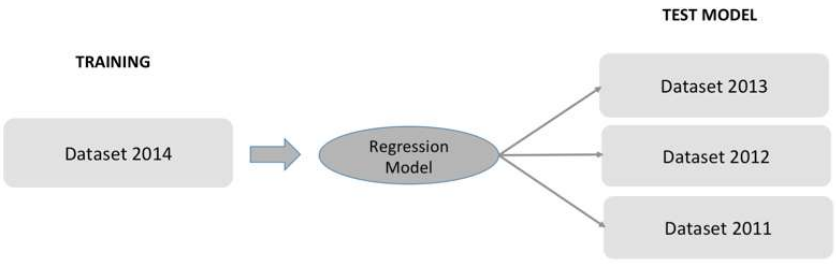

Fig. 2.4 Testing procedure (year 2014)

Each model is tested in 3 different datasets and trained in one dataset. We therefore obtain three MSE per model. We compute the mean MSE for each model. The performance of our preferred model (the named best model) is also tested with an extra dataset, which includes data for 2015. For comparison, the linear model has also been trained using the whole dataset from 2011 to 2014, though it is not recommended to evaluate the model using the training because this method generates overfitted models.

2.3.2 Methods

Considering the No Free Lunch theorems ([64]), we aim to evaluate different methods to disentangle the specific preferred technique according to the testing procedure. More precisely, this paper has tested Multiple Linear Regression, Support Vector Machines and Evolutionary Algorithms. Other methods such as Ridge Regression, Lasso Regression, and Random Forest are also briefly mentioned. All these models are adjusted using the R program language ([47]).

Artificial neural networks (ANNs) are a family of models inspired by biological neural networks and are used to estimate or approximate functions that can depend on many inputs and are generally unknown. This technique has been extensively used for forecasting tasks with some satisfactory results ([68]). Some disadvantages include its black-box nature, greater computational burden, proneness to overfitting, and the empirical nature of model development [62]). A neural network can be thought of as a network of neurons organized

in layers. The predictors or inputs form the bottom layer, and the forecasts or outputs form the top layer. Once we add an intermediate layer with hidden neurons, the neural network becomes non-linear, and according to some test we have carried out, a neural network with hidden layers is necessary in this case to obtain relevant results. However, the model is non-linear and its results are difficult to interpret. For this reason, neural networks have not been further considered in this analysis since we are looking for linear models.

Basic Linear Regression

Linear regression is a simple approach for adjusting a quantitative response Y on a series of regressors X. It assumes that there is a linear relationship between X and Y ([34]). We can write this linear relationship (without time subscripts for simplicity) as:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \tag{2.2}$$

Where $k+1$ coefficients must be estimated, let say $\hat{\beta}_0, \hat{\beta}_1, ...$ and $\hat{\beta}_k$. ε represents the error and contains the variability of the dependent variable not explained by the X. There are several approaches to obtaining the parameters; we have used one of most common approaches that involves minimizing the least squares criterion – ordinary least squares regression ([56]). If $\hat{Y}$ is a vector of n predictions, and Y is the vector of observed values corresponding to the inputs to function which generated the predictions, we choose $\hat{\beta}_0, \hat{\beta}_1, ...$ and $\hat{\beta}_k$ to minimize the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{2.3}$$

With $(\hat{Y})$:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_1 + ... + \hat{\beta}_k \hat{X}_k + \varepsilon \tag{2.4}$$

The values $\hat{\beta}_0$, $\hat{\beta}_1, \ldots,$ $\hat{\beta}_k$ that minimize (2.3) are the multiple least squares regression coefficient estimates ([37]).

Ridge and Lasso Regression

Ridge regression is very similar to least squares, except that the ridge coefficients are estimated by minimizing a slightly different expression. The regression coefficient estimates $\hat{\beta}^R$ are the values that minimize ([32], [34]):

$$RSS + \lambda \sum_{j=1}^{k} \beta_j^2 \tag{2.5}$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately. The second term in expression 2.5 is called a shrinkage penalty. When $\lambda = 0$, the penalty term has no effect on the criteria, and ridge regression will produce least squares estimates. However, as $\lambda \rightarrow \infty$ the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach to zero. Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates (one for every λ). To decide the best λ we need to use either a grid of values or previous analysis using different values and some tests among them.

Least Absolute Shrinkage and Selection Operator (Lasso) regression ([59]) can be a solution when we only want to consider the most important variables. With ridge regression and when as $\lambda \rightarrow \infty$ the coefficient can be close to zero, but none will be set to exactly zero. Sometimes, when the number of variables is too large, this can make it difficult to understand the model. The Lasso alternative forces some estimates to be exactly zero when the parameter λ is large enough. To achieve this, the Lasso coefficients minimize the following expression:

$$RSS + \lambda \sum_{j=1}^{k} |\beta_j| \tag{2.6}$$

In this case, the penalty has the effect of forcing some of the coefficient estimates to be to zero when λ is large enough. Ridge and Lasso regression have been tested in R using *glmnet* R package ([22]). In both Ridge and Lasso regression, several values for the λ hyperparameter have been tested. In particular, a regular sequence of 100 values from 2^{10} to 2^{-2}, and using an increment by $((2^{10} - 2^{-2})/(100-1))$, has been evaluated, chosen for λ the value that minimizes the MSE in the training set for each case. In other words, using the training set, we obtain the coefficients of the model and the MSE value for all the λ alternatives. Finally, the model (coefficients and λ) with the lowest MSE value is selected. A more precise way to do this in order to avoid overfitting, is to employ an additional dataset (validation set) to optimize the hyperparameters, and only use the training set to calculate the coefficients. But in our case we do not have enough data to do this.

Support Vector Machines

Support Vector Machines (SVM) have been widely used for function estimation ([16], [53]), known for our case Support Vector Regression (SVR). SVM is a generalization of a classifier called the maximal margin classifier ([50]) to accommodate non-linear class boundaries.

SVM can be applied not only to classification problems but also to the case of regression. It contains all the key features that characterize maximum margin algorithm: a non-linear function is learned by linear learning machine mapping into a high dimensional kernel induced feature space and, the capacity of the system is controlled by parameters that do not depend on the dimensionality of the feature space. In this paper we have chosen a linear kernel.

In SVM regression, the input x is first mapped onto a m-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space. Using mathematical notation, the linear model (in the feature space) $f(x,w)$ is given by:

$$f(x,w) = \sum_{j=1}^{k} w_j g_j(x) + b \tag{2.7}$$

Where $g_j(x)$ and $j = 1 \ldots k$ denotes a set of nonlinear transformations and b is known as the bias term. The quality of estimation is measured by the loss function $L(y, f(x,w))$. SVR uses a type of loss function called ε-insensitive loss function ([16]):

$$L(y, f(x,w)) = \begin{cases} 0 & if |y - f(x,w)| \leq \varepsilon \\ |y - f(x,w)| - \varepsilon & otherwise \end{cases} \tag{2.8}$$

SVM regression ([11]) performs linear regression in the high-dimension feature space using ε – insensitive loss function and, at the same time, tries to reduce model complexity by minimizing $||w||^2$. SVM regression (SVR) has been adjusted using *e1071* R package ([42]). We use ε-regression to predict one value based on another ([35]). To improve the performance of the support vector regression, we have executed a grid search looking for the best values for ε. There is also a cost parameter which we can change to avoid overfitting. The process of choosing these parameters is called hyperparamether optimization ([5]), or model selection.

Evolutionary Algorithms

The Genetic Algorithm (GA), ([26]) is a method for solving both constrained and unconstrained optimization problems based on natural selection, i.e., the process that drives biological evolution.The genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses these to conceive the children for the next generation. Over successive generations, the population *evolves* towards an optimal solution. The genetic algorithm uses three main types of rules at each step to create the next generation from the current population: i) *selection rules* select the individuals, named parents, which contribute

to the population for the next generation; ii) *crossover rules* combine two parents to conceive offspring for the next generation and iii) *mutation rules* that apply random changes to individual parents to conceive.

However, many real-world problems involve simultaneous optimization of different objectives which are at times hard to evaluate, and often contradictory (otherwise they would be redundant). A Multi-Objective Problem (MOP) can be formulated as:

$$\begin{array}{c} min(f_1(x), f_2(x), ..., f_k(x)) \\ with \quad x = (x_1, x_2, ..., x_n) \end{array} \tag{2.9}$$

In most MOPs, there is no single solution, but a set of equivalent solutions, which is called the Pareto optimal set (or Pareto front). Of all heuristics, Evolutionary Algorithms (EAs) have experienced a great development for their ability to find solutions to complex problems in reasonable time. EAs, as a heuristic tool, avoid some of the problems that traditional optimization methods have had as point estimation and removing regions ([66]). In traditional methods, the execution time rises significantly with increasing dimensionality of the search space; they have troubles with nonlinear problems and with discontinuous functions and they are highly sensitive to numerical precision as it influences the results.

The term EA is used to encompass those heuristic optimization methods that base their operation on the simulation of evolutionary processes of nature mentioned by Darwin ([26]) and operate generating sequences of populations and using mechanisms selection and variation, implemented in the form of genetic operators. A reproduction probability, depending on their quality, is assigned to each solution. The population evolves in time through crosses between individuals and mutations.

The Multi-Objective Evolutionary Algorithms (MOEAs) have been successfully applied to various MOPs, see ([14], [69]) for a comparative study. One of the best known MOEAs is Non-Dominated Sorting Genetic Algorithm (NSGA) ([54]). In the second version (NSGA-II), presented by ([19]), the best individuals in each generation are preserved. This method is elitist, however, and it can sometimes eliminate Pareto optimal solutions and let non-dominated solutions into the current population, although some of them are not Pareto optimal. NSGA-II has been the selected algorithm for our experiments. Although this paper does not deal with a MOP, NSGA-II algorithms are well suited for dealing with mono-objective problems. Our aim using an NSGA-II algorithm is to develop it in future work including more objectives in the optimization process and compare it with previous mentioned methods.

To briefly explain the NSGA-II configuration, we assume the linear model:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon \tag{2.10}$$

We want to estimate coefficients $\hat{\beta}_0$, $\hat{\beta}_1, \ldots,$ $\hat{\beta}_k$ which minimize the following expression:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{2.11}$$

In this case, the accuracy of the estimated model is evaluated through the MSE. $\hat{Y}$ represents a vector of n predictions (for a sample of size n) and Y is the vector of observed values corresponding to the inputs of the function generating the prediction. As all data sets have the same size (365 days), using the MSE instead of the RSS does not produces significant differences. Of course, MSE values are obtained and taken into account for the rest models.

As was explained in the previous section, the model has sixteen variables ($X_1, \ldots, X_{16}$), and seventeen parameters (including the intercept). For this reason, each individual in the population have seventeen genes. Table 2.3 represents an example of a chromosome for the first model previously explained with random values for corresponding $\hat{\beta}_i$. For this example, we consider that we want to estimate the total log market demand at day t ($\hat{Y}$).

Table 2.3 Gene representation (example)

import	**export**	**Jan**	**Feb**	**Mar**	**Ap**	**May**	**Jun**	**Jul**	**Ag**	**Sep**	**Oct**	**Nov**	**Dec**	**labDay**	**LogPrice**	β_0
0.50	0.14	1.50	2.30	0.02	1	0.08	0.10	0.05	0.20	0	0.07	0.08	0.01	0.20	0.05	-3

The main features of the EAs are included in table 2.4 To obtain these values (excluding number of genes), some tests trying out several values have been performed.

Table 2.4 EA parameters

Number of genes	17
Population size	100
Mutation probability	0.20
Crossover probability	0.70
Evaluations	500,000
Executions	15

As we have tested in our case, an increase in population size does not contribute to improve MSE. An EA execution is a stochastic procedure since we run 15 executions in each case to obtain the results, which will be shown hereafter using the parameters of the

preferred model (that with lowest MSE). NSGA-II algorithm has been tested in R using the *mco* package ([61]).

Random Forest

When facing non-linear regression problems, an alternative approach is to sub-divide the space into smaller regions, where the interactions are more manageable. We then partition the sub-divisions again – called recursive partitioning – until finally we obtain chunks of the space which are so slim that they can be fitted with simple models. Thus, the global model has two parts: one is the recursive partition, and the other is a simple model for each partition group. Prediction trees use the tree to represent the recursive partition. Each of the terminal nodes (or leaves) of the tree represents a group of the partition. For regression, the predicted value at a node is the average response variable for all observations in the node.

Using Random Forest ([8]), we can tackle some of the disadvantages of regression trees such as accuracy and instability. The basic algorithm consists of generating a forest of many trees, and then growing each tree on an independent bootstrap sample from the training data. The trees are then averaged to obtain predictions. Although our goal in this paper is to estimate a linear model, we have also tested random forest to obtain insights about the importance of the different variables in this context. Some key facts can be observed from the tree structure. The tree visualization shows us the proximities between variables. These proximities do not just measure similarities between variables, but they also consider their importance. Two cases that have quite different predictors might have large proximity if they differ only in unimportant variables. Two cases that have quite similar values of the predictors might have small proximity if they differ in inputs which are important ([30]). For adjusting Random Forest models, we have used the *RandomForest* R package ([39]).

2.4 Results and testing

In this section, results obtained through different techniques are discussed. For each case the accuracy of the model is evaluated through the MSE. Data comes from OMIE daily markets. We use data from 2011 to 2014 (one dataset per year). We also use data for 2015 to check the estimation results. For each execution, the obtained p-values are shown. Other significance tests have been considered in the case of linear model as the Anova test ([33]) and the Wald test ([63]) for linear models. In summary, two types of findings are going to be presented, coefficients (elasticities) values on the one hand, and, on the other hand, statistical tests for comparing models. Economic and statistical are the two criteria to decide.

2.4.1 Economic results: comparison of elasticities

Our aim is to obtain insights about the contribution of some variables (focusing mainly on prices) in energy demand. Regarding to the dataset, two experiments have been performed:

1. The first one uses data from 2011 to 2014 and, to select a final model among the different alternatives (one per year), the MSE criterion as well as a multiple test set validation procedure have been selected. This testing procedure contributes to avoid overfitting as it has been previously explained.

2. For comparison purposes with the prior case, we have also trained the linear model using the entire dataset from 2011 to 2014 and, in the second case, using only data from year 2015. The obtained models cannot be used to predict demand because they are overfitted models (training and test sets are the same). But, this is not the purpose. These models are only used to analyze similarities with the model provides in the first step. Considering the data for the whole period, we leave the year-specific price elasticity unrestricted. To include this possibility in the specification, we define:

$$LogPriceYear_i = LogPrice * (d_{\mathrm{i}}); where d_i = \left\{ \begin{array}{c} 1 \text{ if year(date)} = i \\ 0 \text{ otherwise} \end{array} \right\} \tag{2.12}$$

Taking all these facts into account, only those models providing results are shown in the next table.

Table 2.5 Regression results

Coefficients with different techniques and samples								
	Multiple test set validation 2011-2014				***Training and testing with the entire dataset 2011-2014***		***Training and Testing with year 2015***	
Variable	**OLS**	**Ridge Reg.**	**Lasso Reg.**	**SVR**	**OLS**		**OLS**	**SVR**
Import	0.02	0.01	0	0	0.02		0.01	0
Export	-0.02	0	0	0	-0.03		-0.02	0
January	0.04	0.01	0	2.45	0.06***		0.07	2.07
February	0.06***	0.01	0	6	0.11***		0.10***	6
March	-0.05***	0	0	2	-0.02		-0.04*	2
April	-0.20***	-0.01	0	-2.08	-0.10***		-0.19***	-2.23
May	-0.10***	0	0	-2.38	-0.12***		-0.10***	-2.59
June	-0.02	0	0	-0.84	-0.06***		-0.03	-0.77
July	0.08***	0	0	-0.66	-0.03***		0.07***	-0.47
August	-0.01	0	0	-3	-0.08***		-0.01	-2.80
September	0.01	0	0	0.53	-0.03***		0.01	0.79
October	-0.05***	0	0	-1	-0.09***		-0.52***	-1
November	0	0	0	-1.17	-0.06***		0	-1
WeekDay	0.15***	0.01	0	7***	0.15***		0.16***	8.10***
Intercept	10.35***	10.19	10.13	9.69	10.30***		10.35***	9.68
Price (log)	-0.08***	-0.02	0	3.44	LogPrice2011 -0.07*** LogPrice2012 -0.05*** LogPrice2013 -0.06*** LogPrice2014 -0.06***		-0.08	3.63
Mean MSE	0.01	0.02	0.02	0.02	0.01		0.02	0
Mean elapsed time (seconds)	0.05	0.06	0.09	21.03	0.01		0.09	0.09
R^2 (training set)	0.77	0.19	0	0.60	0.63		0.76	0.62
*** p-value <0.01 * p-value <0.05								

Both MSE and R^2 values from OLS model slightly outperforms SVM, Ridge and Lasso models. Although, Lasso regression is a method of model building and variable selection that can be applied to many types of regression, including ordinary least squares, logistic regression, and so on, in this case, it is not a satisfactory technique for our purposes, as it does not provide us any distinguished weight to the variables. Lasso is also used to carry out variable selection. The variable selection objective is to recover the correct set of variables that generate the data or at least the best approximation given the candidate variables. Lasso rely on assumptions in order to work. The first is sparsity, (only a small number of variables may be relevant), and the second one is that the irrepresentable condition must hold (the relevant variable may not be very correlated with the irrelevant variables). According to table 2.2, the irrepresentable condition is satisfied. However, the used model has a very limited number of variables and it appears that the model is not able to distinguish any special relevant variables.

Very similar results are found using Ridge model. In both cases, all coefficient values are zero or very close to zero, with the minor R^2 values. Both models penalize the magnitude of coefficients, doing that, they ignore nonsignificant variables that may, nevertheless, be interesting or relevant to obtain the best fit. For this reason, these models have not been included in the execution analyzing the other cases. In relation to the significance test for the linear model and in this first case (multiple test set validation), Anova test finds some additional significance coefficients, maintaining the same conclusions for the rest. It also finds strong significance (0.001 as reference value) for the coefficient of import, while the coefficient of export is still significant at standard levels. They show the expected sings since demand increase with import and decrease with export (they are expected because when demand exceeds capacity, the country needs to import while the reverse is true when production exceeds demand). The Wald Test confirms exactly the p-values conclusions. In all relevant cases, we find that an increase in prices negatively affects demand. All the models confirm this result but with very low values.

For the second case, only OLS, SVM and NSGA-II have been tested (according to the previous results provided by Lasso or Ridge model in the first case). Using data from 2011 to 2014 as the training set, SVR has not been able to generate any model whatsoever. As has been explained in previous sections, we use ε-regression to predict one value based on another. This model, and especially the model implementation using *e1071* R package, presents difficulties to find a suitable model when the variable to predict has a very low variance (as shown in the next table).

Table 2.6 Standard deviations of variables for the period 2011 – 2014

Variable	**SD**
LogPrice	0.67
LogMarketDemand	0.13
Hydro	0.07
Nuclear	0.03
Coal	0.09
ComCycle	0.07
FuelGas	0.01
RegEsp	0.12
Import	0.01

OLS Model obtained using data from 2011-2014 as training (and test) set, obtain the lower MSE value because of the overfitting effect. But it provides information about the prices elasticities, obtaining very similar and negative values for price coefficients for the years 2012, 2013 and 2014, and a slightly different value for year 2011. OLS model using data from year 2015 also obtain a negative coefficient for price. SVR obtain a very different value for all the coefficients with a lower MSE and R^2, and with only one coefficient statistically significant.

Regarding to NSGA-II model, this algorithm does not generate suitable results either, because in both cases (1 and 2), it provides a unique value for the coefficients (1), with, i.e., 3.5 as MSE value for the first case, and WeekDay and price coefficients as significant ($p-value < 0.01$).

The Random Forest algorithm has been used to obtain additional information about the importance of every variable. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. Each time a variable is used to split a node, the Gini coefficient for the child nodes are calculated and compared to that of the original node. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). For regression, it is measured by residual sum of squares. Variables that result in nodes with higher importance have a higher value in Gini coefficient.

Regarding the relation between market demand and price - Gini coefficient indicates that, generally and by far, the first variable in importance is the price to predict market demand. The second variable in importance, and with a value away from less important variables, is the *WeekDay* variable (indicating whether it is a working day or not). Months from January to December have slightly similar values. According to Gini coefficients, the least important

variable is the *exported* variable, indicating whether or not electricity has been exported on that given day.

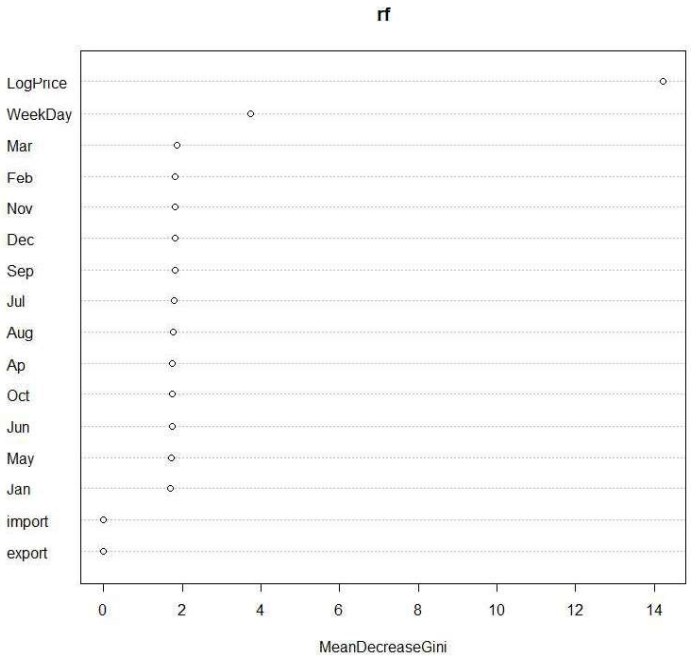

Fig. 2.5 Variable importance by Random Forest

2.4.2 Checking the model from a statistical point of view

To build a linear model, the model should conform with some assumptions of linear regression. Some of these assumptions have been checked during the execution process. Concretely:

- Assumption 1: residual mean equals 0.
- Assumption 2: no autocorrelation of residuals. If the null hypothesis is rejected means that there is a definite pattern in the residuals.
- Assumption 3: the main predictor variable and residuals are uncorrelated. If the p-value is lower than 0.05, null hypothesis that true correlation is 0 can be rejected.

Although we have chosen the best model based on the MSE results, other measures have been introduced to give us information about the model accuracy. These measures are R^2, adjusted R^2, F-statistic, the Akaike's Information Criterion (AIC) ([1]), the Bayesian

Information Criterion (BIC) ([51]), the Mean Absolute Percentage Error (MAPE) and the MinMaxAccuracy. We express these last two criteria (while the remaining are very well-known) as:

$$MAPE = mean(\frac{abs(\hat{y}_i - y_i)}{y_i}) \tag{2.13}$$

$$MinMaxAccuracy = mean(\frac{min(y,\ \hat{y})}{max(y,\ \hat{y})}) \tag{2.14}$$

$min(y,\ \hat{y})$ represents the min value for each par (actual_value, predicted_value). In the case of MAPE, the lower the value, the better the specification, while the reverse is true for MinMaxAccuracy.

Considering the definitions explained in the last section, and the model explained by equation 1 with results in table 2.5, we provide additional values to measure the goodness of fit in table 2.7.

Table 2.7 Model fit evaluation (predicting MIBEL demand taking logs)

	Fit measures						
	Multiple test set validation 2011 - 2014				**Training and testing with the entire dataset 2011 - 2014**	**Training and Testing with year 2015**	
Variable	**OLS**	**Ridge Reg.**	**Lasso Reg.**	**SVR**	**OLS**	**OLS**	**SVR**
R^2_{adj} (higher the better)	0.27	0.15	<0	<0	0.62	0.75	<0
f – Statistic (higher the better)	9.50	4.97	0	13.55	127.70	74.10	365
AIC (lower the better)	-569	-466	-391	-409	-3155	1975	1938
BIC (lower the better)	-585	-483	-408	-410	-3044	1959	1937
MAPE (lower the better)	0.01	0.01	0.01	0.01	0.01	0.25	0.25
MinMaxAccuracy (higher the better)	0.99	0.99	0.99	0.99	0.99	0.75	0.75
Linear Model Assumption 1	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected
Linear Model Assumption 2	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected
Linear Model Assumption 3	Not rejected	Rejected	Rejected	Rejected	Rejected	Rejected	Rejected

Table **??**, in general, confirms the results shown in table 2.6, in the sense that it does not allow to classify the performance of the algorithms in a very different way. Considering all measures, it appears that OLS models slightly outperforms the rest.

2.4.3 An additional experiment to robustness

An additional question regarding to the algorithm results, is answered in this section; it refers to what would occur if the data had been different. With this purpose in mind, fake data has been produced. The procedure includes generating a random vector of coefficients using uniform distribution. Then, by composing different values, new numbers for energy demand were obtained for each year and dataset. In summary, if we have the equation:

$$x_t = \alpha + \beta y_t + \gamma' z_t + u_t \tag{2.15}$$

where x represents the log-price of electricity; α, β and γ are randomly generated parameters, z stands for the dummies variables (zeros o ones), fake values are obtained with this expression:

$$y\_fake_t = \alpha + \beta f(x_t) + \gamma' z_t + u_t \tag{2.33}$$

where *f(x)* can be one of the following:

1. Identity function (*f(x)* = *x*)
2. $f(x) = x^2$
3. $f(x) = e^x$
4. $f(x) = \frac{1}{1+e^{-x}}$

Using this generated data, OLS, SVR and NSGA-II model are training to fit the fake data using the procedures explained in the previous sections. The results show that regardless of the method followed to generate the α, β and γ coefficients, and the expression used for *f(x)* , the linear model with OLS obtain values close to one for R^2 and $R^2{}_{adj}$ (the procedure explained in 2.3.1 has been again applied here to avoid overfitting). This is a very good fit that is only achieved by the NSGA-II algorithms with a small difference. Only one exception can be mentioned, considering $f(x) = x^2$, the best fit is, again, achieved by the linear model and, now, the results of the NSGA-II algorithm are far from this performance. Another point is that the linear model using OLS finds most of the coefficients to be significant. These results are consistent with the method used to obtain the 'fake' values for energy demand. As linearity is implicitly into the model, the linear model obtains the better fit.

Table 2.8 Model fit evaluation for 'fake' data

Fit measures												
	Linear			Polynomical			Exponential			Sigmoid		
Variable	**OLS**	**SVR**	**NSGA-II**	**OLS**	**SVR**	**NSGA-II**	**OLS**	**SVR**	**NSGA-II**	**OLS**	**SVR**	**NSGA-II**
MSE	0	8.75	20	4.5	20.40	142.65	50.50	92	1266.89	0	8.61	8.35
R^2	1	<0	<0	0.73	<0	<0	0.18	<0	<0	1	0.98	<0
R^2_{adj} (higher the better)	1	<0	<0	0.72	<0	<0	0.14	<0	<0	1	<0	<0
f – Statistic (higher the better)	0	<0	<0	139	<0	<0	8.46	63	<0	185778	<0	<0
AIC (lower the better)	-20951	1523	2136	1298	1933	2862	2475	2608	3663	-1584	1499	1825
BIC (lower the better)	-20967	1522	2120	1282	1932	2846	2459	2607	3647	-1600	1498	1808
MAPE (lower the better)	0	0.21	0.31	0.04	0.19	0.57	0.16	0.23	0.78	0.002	0.25	0.25
MinMaxAccuracy (higher the better)	1	0.85	0.70	0.96	0.86	0.44	0.88	0.85	0.21	0.99	0.83	0.78

In this context and according to the results shown in the last table, it appears than a linear and parsimonious model estimated by OLS, despite its apparent simplicity, outperforms the rest. From our point of view, we show strong results to favor simple models for estimating energy demand models.

2.5 Conclusions and policy implications

The paper aims to analyze the evolution of Spanish electricity market from 2011 to 2014 and to contribute to a better understanding of the behavior of the market. We estimate price elasticities as key factor to understand how public policies or economic conditions impact the market structure. One contribution of the paper is the comparison of the elasticity values estimated under very different econometric methods.

The results focused on MSE and R^2 values, show that OLS model slightly outperforms SVM, Ridge and Lasso models. Lasso and Ridge regression do not provide us any distinguished weight to the variables. In both cases, all coefficient values are zero or very close to zero. SVM obtains a worst fit than OLS in most cases. In addition, the chosen SVM implementation presents some difficulties to obtain a model in cases when the variable to predict has a very low variance. Regarding to the tested MOEA, this algorithm does not generate suitable results because it provides a unique value for the coefficients (1), falling into a local optimum.

To about overfitting, the methodology uses, a multiple test set procedure, using different datasets for training and testing. This method, jointly with significance test, helps to assure the robustness of the results.

Focused on the economic results, we find very low reaction of consumers to price movements and we have given robustness to our results through different methods and tests. The results evidence strong significance for the coefficient of import, while the coefficient of export is still significant at standard levels. These values are expected because when demand exceeds capacity, the country needs to import while the reverse is true when production exceeds demand. In all relevant cases, we find that an increase in prices negatively affects demand. The Random Forest algorithm has been used to obtain additional information about the importance of every variable. The first variable in importance is the price to predict market demand. The second variable in importance, and with a value away from less important variables, is the WeekDay variable (indicating whether it is a working day or not). Months from January to December have slightly similar values. According to Gini coefficients, the least important variable is the "exported" variable, indicating if electricity has been exported on that given day.

Finally, an additional experiment has been performed to analyze what would occur if the data had been different. Fake data for energy demand has been produced using linear models. Again, a linear and model estimated by OLS outperforms the rest, showing that simple models are suitable for estimating parsimonious energy demand models.

Chapter 3

Predicting Energy Demand in Spain and Compliance with the Greenhouse Gas Emissions Agreements

3.1 Introduction

This paper uses a very simple model to predict energy demand in Spain for the 2020 scenario. We based our prediction solely on climatic and some variables proxying the economic activity. Our main research question is whether Spain can achieve an energy-mix able to meet the EU commitments by 2020. In general, forecasting the likely path of greenhouse gas emissions is essential to understanding the range of possible effects of climate change. The European Commission (EC) and EU governments agreed on the target of cutting greenhouse gases by at least 20% by 2020 [10], compared with 1990 levels. Hence, it is mandatory for all EU member countries to ensure that between 20 and 30 percent of the consumed energy comes from clean renewable energy sources. This European action on climate change has its antecedents in Articles 17, 18 and 19 of the Directive EU 2009/28/CE of the European Parliament and Council of April 23, 2009, which was transferred to Spain by the RD 1597/2011 of November 4, 2011. The United Nations Climate Change Conference (Paris, December 2015) ratified these agreements. They have been later approved by the European Union Parliament on October 4, 2016.

The methodology used for forecasting energy demand is manifold. Our model is parsimonious since we use a specification considering climatic variables (Heating Degree Days or HDD, Cooling Degree Days or CDD, and volume of rainfall), and economic variables

(activity level proxied only by the Industrial Production Index or IPI)[1]. The IPI measures the monthly development of industrial activity, including extractive, manufacturing, and production and distribution of electricity, water and gas. This indicator reflects the joint development of quantity and quality, independent of the influence of prices. The Instituto Nacional de Estadística (INE) builds the IPI through a survey concerning details of the production of activity branches compiling monthly data for more than 11,500 establishments. According to [48], the industrial sector is the largest consumer of electricity, close to 30% of the total amount.

A second objective of this paper is to compare different regression methods with prediction purposes. We use a Mean Square Error (MSE) criterion to test Linear Regression, Support Vector Machines for Regression (SVR) and Deep Learning Neural Networks. Deep Learning uses machine learning algorithms in order to model high-level abstractions with multiple non-linear transformations. In addition, for greater accuracy and sensitivity in the evaluation, we divide the original data into a training set and a test set, as is explained in section 3. The results show that models based on neural networks significantly improve the MSE criterion when compared to Linear Regression or SVR. On the other hand, if we consider a scenario for the foreseeable future consisting of an increase of IPI similar to that given in previous reporting periods (from November 2008 to December 2011), the simulations predict an energy demand in December 2019 close to 23 thousand Gigawatt hours (GWh). According to our data, this demand will contribute to more than 6 million tons of CO_2 emissions to the atmosphere. Considering historical data from Red Eléctrica de España (REE) and the afore mentioned energy demand there is clear evidence that this energy-mix will allow Spain to meet EU clean renewable energy agreements and, that this will cover between 20% and 30% of total demand.

The remainder of the chapter is organized as follows: Section 3.2 introduces and describes the data used for the empirical exercise; In section 3.3, we explain the different methods and the results obtained; Section 3.4 shows the results of the simulation of energy demand at the end of 2019; Section 3.5 concludes.

3.2 The data

Our aim is to perform the analysis using a parsimonious model fed by as little information as possible. This study only uses climatic variables, and a proxy for industrial production. The model used has been tested against unrestricted models based on demand specifications and

[1]We try another economic indicators but since the industry is the largest consumer of energy, we believe our parsimonious model can fit better and cover our prediction purposes.

this is our preferred model for prediction purposes based on a battery of tests. According to a study published by the BBVA Foundation [40], industry, historically, transfers purchasing power to other sectors. Company profits evolve in line with a yearly exchange rate which is linked to the IPI. Economic crashes are likely to be reflected significantly in this index. The Spanish IPI (adjusted seasonally) reached its lowest value since 2007 in April 2012, with an accumulated depreciation close to 30% during this period [58]. Furthermore, annual series with mean monthly IPI values for the period 2007 - 2014, are highly correlated with average annual expenditure of Spanish households during that time interval. It is therefore assumed that the IPI provides an accurate proxy of the economic activity to be used for estimating different scenarios for economic growth in Spain.

In addition we use climatic variables. Temperature data have been obtained from several sources, including Agencia Estatal de Meteorología (AEMET) weather stations, Ministry of Agriculture, as well as data from Red Eléctrica Española (REE) and from the National Climatic Data Center [44]. Weather stations are located in different areas of Spain: the north; Cantabrian coast; the Meseta Central; the south and Mediterranean Coast. Matching this data and taking daily averages, we built a dataset with daily observations for the period March 1, 2007 to December 31, 2015, with estimated maximum, minimum and average daily temperature. We obtain daily rainfall (PREC) in the same way. Data from the islands (Balearic Islands and the Canary Islands), have not been considered in this paper. Based on these temperatures, HDD and CDD have been calculated using the following formulas [49], using the first one that matches:

$$HDD_t = \begin{cases} 0 & t_{min_t} > t_{base_t} \\ \frac{t_{base_t} - t_{min_t}}{4} & \frac{t_{max_t} + t_{min_t}}{2} > t_{base_t} \\ \frac{t_{base_t} - t_{min_t}}{2} - \frac{t_{max_t} - t_{base_t}}{4} & t_{max_t} \geq t_{base_t} \\ t_{base_t} - \frac{t_{max_t} + t_{min_t}}{2} & t_{max_t} < t_{base_t} \end{cases} \tag{3.1}$$

$$CDD_t = \begin{cases} 0 & t_{max_t} < t_{base_t} \\ \frac{t_{max_t} - t_{base_t}}{4} & \frac{t_{max_t} + t_{min_t}}{2} < t_{base_t} \\ \frac{t_{max_t} - t_{base_t}}{2} - \frac{t_{base_t} - t_{min_t}}{4} & t_{min_t} \leq t_{base_t} \\ \frac{t_{max_t} + t_{min_t}}{2} - t_{base_t} & t_{min_t} > t_{base_t} \end{cases} \tag{3.2}$$

where t represents the day. The results of figure 3.1 are taken as base temperature for calculating HDD and CDD the value of 15.5 degree Celsius. In order to match climatic data to economic data we take mean monthly values.

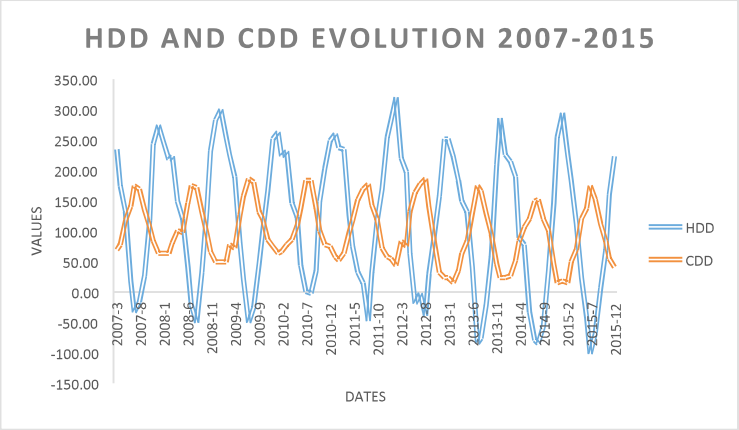

Fig. 3.1 HDD and CDD evolution over the period 2007 - 2015

We also need information about economic activity and prices. Several possibilities are available such as the Gross Domestic Product (GDP), unemployment data, energy prices, Consumer Price Index, etc. We have decided to use the Industrial Production Index (IPI), a monthly time series collected by INE. Available data are shown in figure 3.2.

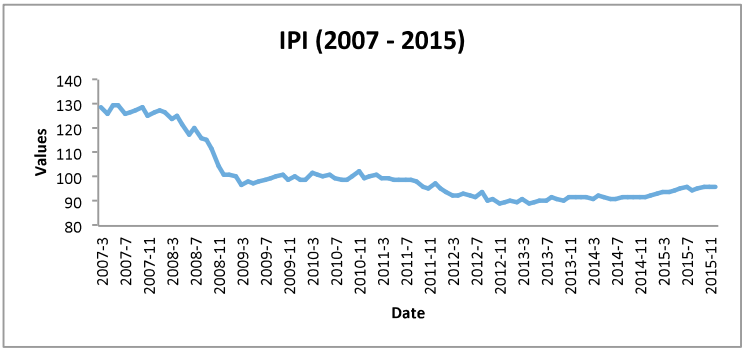

Fig. 3.2 IPI evolution for the period 2007 - 2015

For the purpose of comparison, quarterly GDP growth and unemployment data for the period considered in the analysis are shown in figure 3.3. This data are produced quarterly by the INE.

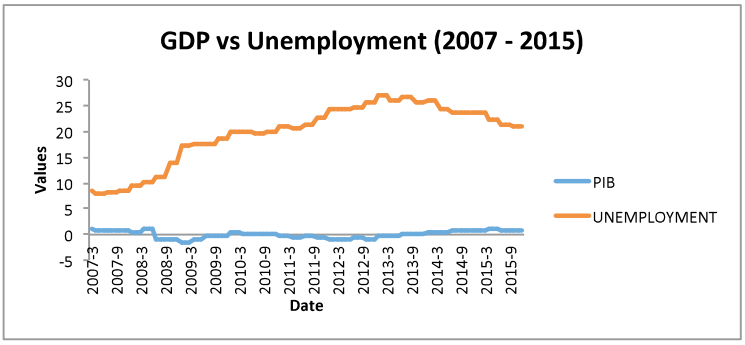

Fig. 3.3 GPD and unemployment evolution (2007 - 2015)

Figure 3.4 shows GDP adjusted taking into account unemployment.

The industrial sector is the largest consumer of electricity (30%), while the services sector accounts for 13% of consumption ([48]).

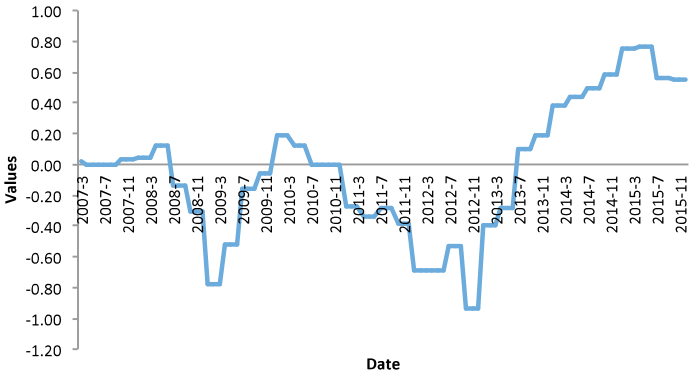

Fig. 3.4 GDP adjusted for unemployment (2007-2015)

Finally, monthly energy demand data in GWh. (from REE and the Ministry of Industry and Energy) are shown in figure 3.5.

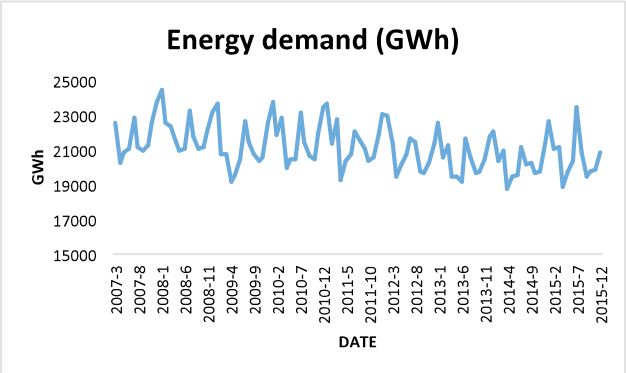

Fig. 3.5 Energy demand in Spain

As we observe all these figures, it is difficult non-parametrically to decide upon the best proxy for energy demand. We can assume that a model which rationalizes the behavior of economic agents, and whose specification includes a proxy for income (GDP) and proxy for prices of energy could provide an adequate alternative, i.e., a proper model of demand. However, our main aim is to provide a parsimonious model to obtain the best possible prediction.

3.3 Methodology and results

In this section we explain the different methods to adjust energy models, while also processing and showing the results. As mentioned previously, our goal is to predict energy demand (GWh.)[2] in Spain to test whether EU commitments can be achieved. The EU agreement forces countries to use between 20% and 30% of total energy using renewable or clean sources. We like to test whether parsimonious models help us in making accurate predictions of energy demand or energy consumption by only using climatic variables and the IPI index. [38] shows that the demand for energy is absolutely inelastic with respect to the price for Spain in the considered period.

3.3.1 Estimation methods

Taking into account the no free lunch theorems [64], we chose to evaluate different methods in order to disentangle the particular preferred technique according to the testing procedure. To be more precise, in this paper we use Linear Regression, Support Vector Machines and a Deep Learning algorithm. All these models are implemented using the R Software [47].

Linear Regression

Linear regression is a simple approach for predicting a quantitative response Y on the basis of a single regression variable X. It assumes that there is a linear relationship between X and Y [34]. We can write this linear relationship as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.3}$$

Where β_0 and β_1 are two parameters that represent the intercept and slope. ε represents the error and contains the variability of the dependent variable not explained by the X.

[2]We denote energy models as we cannot characterize them as demand or supply models. In any case, we acknowledge our interest in predicting energy consumption.

The regression coefficients β_0 and β_1 are unknown, and they are estimated on a sample ($\widehat{\beta_0}$ and $\widehat{\beta_1}$). With these estimated coefficients, we can obtain predictions ($\widehat{Y}$) as follows:

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X \tag{3.4}$$

There are several approaches to obtain the parameters, one of most common approaches involves minimizing a Least Squares Criterion (LSC) ([56]). Ordinary Least Squares (OLS) with heteroskedasticity-autocorrelation robust standard errors ([31]) has been selected for this work. If $\widehat{Y}$ is a vector of T predictions and Y is the vector of observed values corresponding to the inputs to the function which generated the predictions, we choose $\widehat{\beta_0}$ and $\widehat{\beta_1}$ to minimize the residual sum of squares (RSS):

$$RSS = \sum_{1}^{T} (Y_i - \widehat{Y_i})^2 \tag{3.5}$$

In practice, we often have more than one predictor, so Multiple Linear Regression (MLR) is used. Suppose that we have k different predictors, the MLR model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \tag{3.6}$$

In this case, k coefficients have to be estimated, let say $\widehat{\beta_1}$, $\widehat{\beta_2}\ldots$ and $\widehat{\beta_k}$. In this case, we can obtain predictions ($\widehat{Y}$) as follows:

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \widehat{\beta_3} X_3 + ... + \widehat{\beta_k} X_k \tag{3.7}$$

The values $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}\ldots$ and $\widehat{\beta_k}$ that minimize 3.5 are the multiple (ordinary in our case) least squares regression coefficient estimates ([37])

Support Vector Machines

Support Vector Machines (SVM) has been widely used for function estimation ([16], [53]), known for our case Support Vector Regression (SVR). SVM is a generalization of a classifier called the maximal margin classifier ([50]) in order to accommodate non-linear class boundaries. SVM can be applied not only to classification problems but also to the case of regression. It contains all the main features that characterize maximum margin algorithm: a non-linear function is learned by linear learning machine mapping into a high dimensional kernel induced feature space and, the capacity of the system is controlled by parameters that do not depend on the dimensionality of the feature space. In SVR, the input X is first mapped onto a m-dimensional feature space using some fixed (nonlinear) mapping, and then a linear

model is constructed in this feature space. Using mathematical notation, the linear model (in the feature space) $f(X,w)$ is given by:

$$f(X,w) = \sum_{1}^{k} w_i g_i(X) + b \tag{3.8}$$

Where $g_i(X)$ and $i = 1...k$ denotes a set of nonlinear transformations. b is known as the bias term. The quality of estimation is measured by the loss function $L(y, f(X,w))$. SVR uses a type of loss function called ε – insensitive loss function ([16]):

$$L(y, f(X,w)) = \begin{cases} 0 & |y - f(X,w)| \leq \varepsilon \\ y - f(X,w)| - \varepsilon & otherwise \end{cases} \tag{3.9}$$

In addition to use an ε – insensitive loss function, SVR tries to reduce model complexity by minimizing $||w||^2$ (see [11], for additional details). SVR has been adjusted using *e1071* R package ([42]). In our work, we have use a linear kernel and epsilon support vector regression function.

By default, ε takes value of 0.1. But in order to improve the performance of the support vector regression we have executed a grid search looking for the best value for ε. There is also a cost parameter which we can change to avoid overfitting. The process of choosing these parameters is called hyperparameter optimization ([5]), or model selection. We do not have enough data to consider an extra validation set to caliber these parameters.

Deep Learning Neural Networks

Some techniques try to replicate the efficiency and robustness by which the human brain represents information and obtains knowledge. These works motivated the emergence of the subfield of deep machine learning, which focuses on computational models for information representation that exhibit similar characteristics to the neocortex ([4]). Artificial Neural Networks (ANNs) are a family of deep learning models inspired by biological neural networks and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. ANNs are generally presented as systems of interconnected neurons that exchange messages among them. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. This technique has been extensively used for forecasting tasks with some good results ([68]). As disadvantages we can quote its "black box" nature, its greater computational burden, its proneness to over-fitting, and the empirical nature of model development. A neural network can be thought of as a network of "neurons" organized in layers. The predictors or input

form the bottom layer, and the forecasts or output form the top layer. Once we add and intermediate layer with hidden neurons, the neural network becomes non-linear. Configuring the neural network, activating function, layers, etc., are not trivial tasks.

3.3.2 Empirical results

For accurate assessment, we randomly divide the original data into two sets (getting each set data from all the months), a training set and a test set. The test set contains data from January 2008 to June 2008 and from July 2014 to December 2014. The remaining data is used for training. We have randomly selected different months and years to insert some variability in the data.

As explained previously, our main objective is to estimate the parameters of the model represented by the following equation:

$$ENERGY\_DEMAND = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 IPI \tag{3.10}$$

The following figures display results graphically, in each case, we first show predicted values using test set, and subsequently, using training set. The following 2 figures refer to results using linear regression with ordinary least squares method:

Fig. 3.6 Results on test set using linear regression

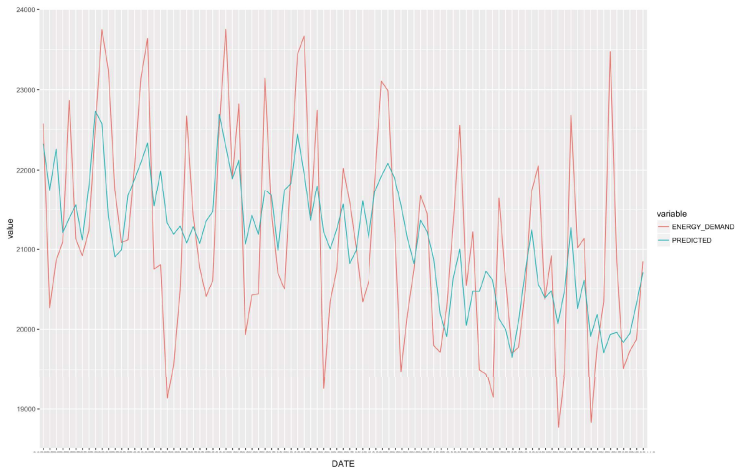

Fig. 3.7 Results on training set using linear regression

Using SVM, we obtain the following graphics:

Fig. 3.8 Results on test set using SVR

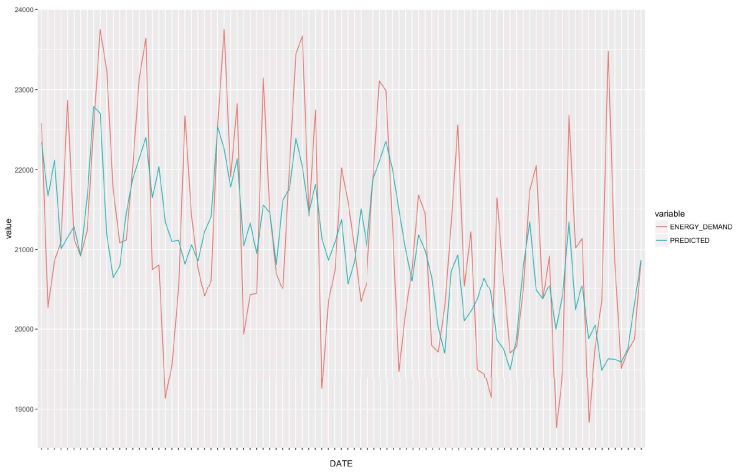

Fig. 3.9 Results on training set using SVR

As we show in the next section, SVR error (MSE) exceeds that of the classical regression MSE.

In the case of Deep Learning Neural Networks, the architecture we have employed consists of two hidden layers with the same number of neurons. The neural network is trained with stochastic gradient descent using back-propagation. This architecture has been tested with 20, 50, 100 and 200 neurons in each layer. The results are shown Table 3.1. These results consider a base temperature value of 15.5.

Table 3.1 Deep Learning Neural Network experiments

Neurons in layer	MSE in test	MSE in training
20	1,168,439	256.4161
50	346,196.6	365.0929
100	385,315	125.3764
200	479,269.1	229.5746

Table 3.1 shows a known problem of over-fitting exhibited by this method, which may occur when using 100 and 200 neurons per layer. In these cases, the MSE using the training set drops significantly compared with those obtained by other architectures. It becomes harder to avoid over-fitting with small-data and sometimes simpler models are more appropriate. However, the MSE using the test set is greater compared to those obtained using 50 neurons per layer. Figure 3.10 shows prediction results using the training set and 50 neurons in each layer (using a base temperature value of 15.5). The comparison of observed and predicted values in 3.10 indicates that the model adjusted through Deep Learning Neural Networks shows a rather high accuracy.

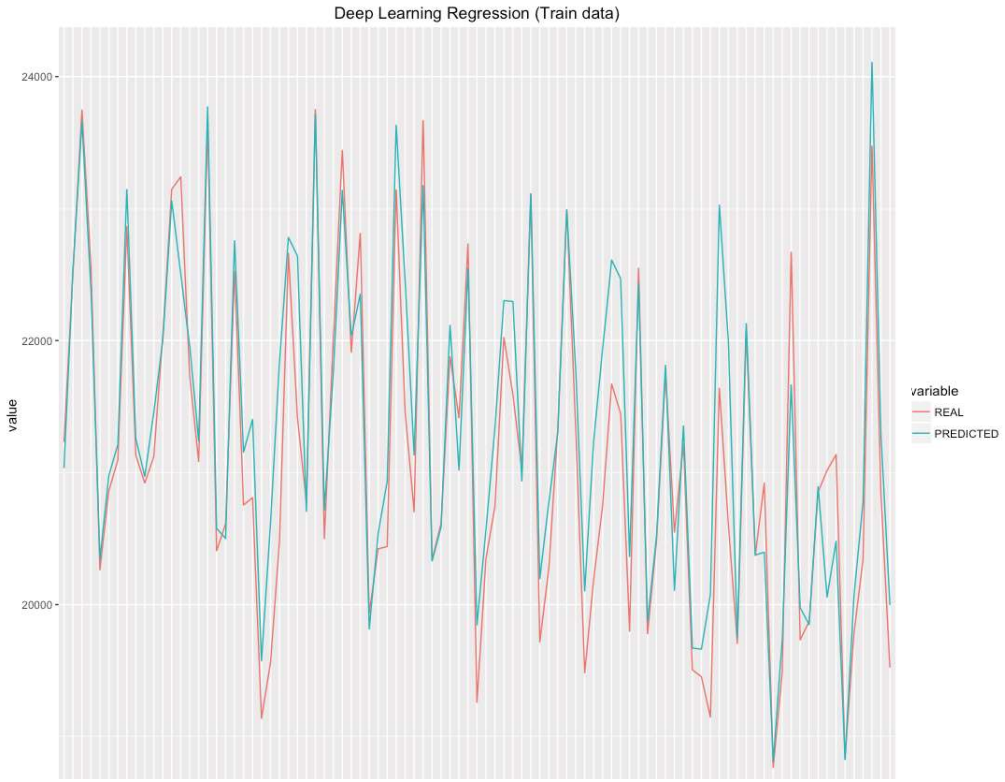

Fig. 3.10 Deep Learning Neural Networks applied to training set

MSE using the training data is lower than using a simpler architecture, but it is significantly higher when the test sample is considered. Therefore, we will work with 50 neurons in each layer according to the tests reported in table 3.1. Results using this architecture are shown in figure 3.11:

Fig. 3.11 Results on test set using Deep Learning Neural Networks

According to the graph, it may seem that this technique obtains a better approach. This feeling is confirmed by the comparison we will carry out next.

Comparison of different methods

To compare the results obtained by different techniques, we show in Table 3.2 the MSE on the test sample taking into account each technique and a base temperature. This comparison of the three procedures used points toward the use of deep learning results for predicting total energy demand. As we have previously explained, Linear Regression uses Ordinary Least Squares Criterion.

Table 3.2 MSE of different methods

Base Temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	973,182.1	1,027,007	511,557.6
15.5	970,816	1,017,013	346,196.6
16.5	968,461.8	1,007,333	358,933.9

We can also show results using the coefficient of determination (R^2). An R^2 of 1 indicates that the regression line perfectly fits the data. Table 3.3 illustrates that the deep learning technique obtains the best value.

Table 3.3 R^2 of different methods

Base Temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.439	0.392	0.748
15.5	0.440	0.392	0.801
16.5	0.442	0.374	0.708

Coefficients for HDD, CDD, and intercept parameter are significant (p-value less than 0.001). Tests do not find significance for PREC (precipitations) or IPI coefficients. Despite the fact that industry sector is the largest consumer of electrical energy taking into account other productive sectors such as services (it has been mentioned in 3.2), models have not found significance for the IPI coefficients. We tried again using GDP (quarterly GDP growth in percentage) instead of IPI. In this case, and using linear model we found little significance for GDP coefficient (p-value between 0.1 and 0.05, and no find significance using other models). But we obtain worse results for MSE and R^2 in all cases and with all the techniques. For example, with a base temperature value of 14.5 the linear model using IPI obtains a MSE

value in test of 973,182.1, but using the GDP as input variable instead of the IPI, the obtained value is 1,039,117. For this reason, we have chosen not to include GDP in this model.

Regarding the execution time in seconds required for each technique, the linear model obtains the results in 0.0017 seconds, the SVR lasts 55 seconds and deep learning methods take 2.8 minutes (on a laptop with 2,9 GHz Intel Core i5 and 8 GBs of RAM). We can conclude that for the problem at hand the lower values of the MSE criterion compensates for the longer execution time.

Tuning SVR parameters implies to increase execution time from 0.65 seconds to 46. Grid search sets the ε value to 0.1 using 14.5 as base temperature, while 0.8 and 0.7 have been the selected values for 15.5 and 16.5 base temperature values respectively.

Figure 3.12 shows results looking for ε best values, taking into account cost optimization and a base temperature value of 16.5.

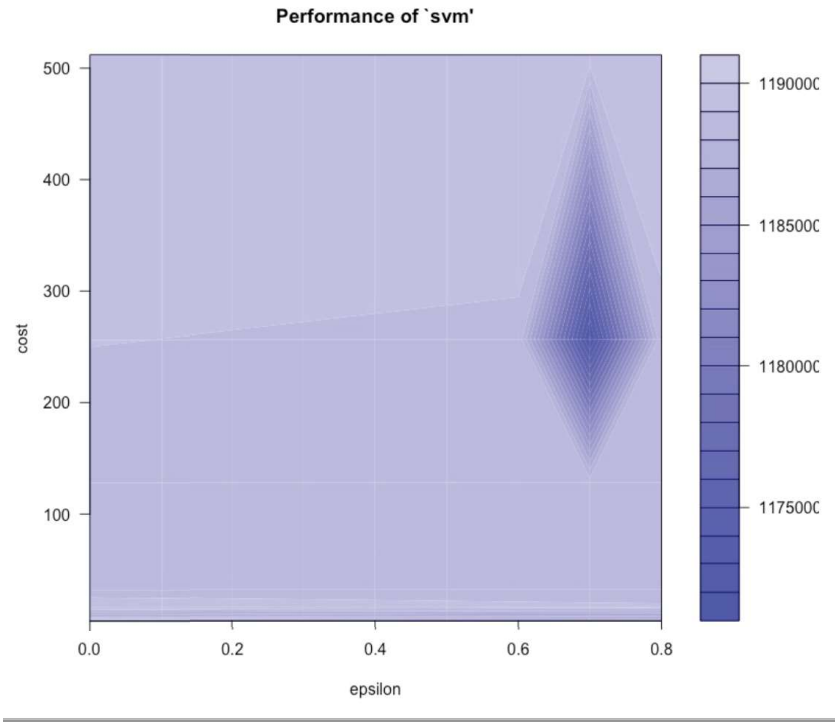

Fig. 3.12 Tuning ε value for SVR and base temperature 16.5

In addition to comparing the different methods, we validate our model doing an analysis of variance (ANOVA) of monthly residuals. We do not detect evidence of heteroscedasticity (p-value is 0.19 greater than the reference value 0.01). However, when we are working with aggregated monthly data, the volume of information is insufficient for a given year. If we

join residuals from all the time period, the ANOVA shows evidence of significant differences between the prediction errors generated for each technique.

3.3.3 Adjusting a rational demand model

The previous model has not properly derived from any optimization problem taking into account behavior of economic agents. We postulate here a model that takes into account energy demand as a function of income and prices. (see, for instance, [18]). Some readers might find it strange that in the linear model, IPI coefficient shows non statistical significance. In this context, we have tried other models to improve the MSE and R^2 results. The main difference is to consider the GDP deflator. The GDP deflator is a measure of price inflation/deflation with respect to a specific base year. Taking the nominal GDP (in millions of euros) and the GDP deflator we obtain the real GDP, we express it in logs. Now, to estimate energy demand we use the following equation:

$$ENERGY\_DEMAND = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 log(\frac{nominalGDP}{GDPdeflator}) \quad (3.11)$$

The obtained MSE and R^2 values have been the following:

Table 3.4 MSE using nominal GDP and GDP deflator

Base Temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	1,120,323	1,084,567	2,849,661
15.5	1,118,067	1,081,096	1,238,948
16.5	1,115,824	1,077,402	1,979,832

We can also show results using the coefficient of determination (R^2).

Table 3.5 R^2 using nominal GDP and GDP deflator

Base Temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.35	0.37	Not Applicable
15.5	0.36	0.38	0.29
16.5	0.36	0.38	Not Applicable

Analyzing significance for coefficient values, we can conclude that coefficients for HDD, CDD, and intercept parameter are significant (p-value less than 0.001). Again, tests do not find significance for PREC (precipitations) or $\frac{nominalGDP}{GDPdeflator}$ coefficients.

We have also evaluate results applying natural logs to energy demand as next equation shows:

$$log(ENERGY\_DEMAND) = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 log(\frac{nominalGDP}{GDPdeflator}) \tag{3.12}$$

In this case, R^2 results (we cannot compare MSE values using different predicting values) are shown in table 3.6.

Table 3.6 R^2 using nominal GDP and GDP deflator - taking logs in energy demand

Base Temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.35	0.27	Not Applicable
15.5	0.36	0.36	Not Applicable
16.5	0.36	0.32	Not Applicable

Analyzing these results we can conclude that inserting in the equation the real GDP does not improve the predictive power of the model.

We can now add some extra information to the model taking into account the energy prices (EP - index with base = 100) and the CPI (monthly). The goal is to evaluate is this extra information is useful to improve the prediction. The equation **??** models this case:

$$ENERGY\_DEMAND = \beta_0 + \beta_1 HDD + \beta_2 CDD + \beta_3 PREC + \beta_4 log(\frac{nominalGDP}{GDPdeflator}) + \beta_5 log(\frac{EP}{CPI}) \tag{3.13}$$

Tables 3.6 and 3.7 show the results. This model does not improve the first model MSE or R^2.

Table 3.7 MSE adding information about energy prices and CPI

Base Temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	1,141,071	1,257,050	867,865.4
15.5	1,138,936	1,236,390	1,773,467
16.6	1,136,809	1,222,573	1,953,979

Table 3.8 R^2 adding information about energy prices and CPI

Base Temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.34	0.27	0.34
15.5	0.34	0.29	Not Applicable
16.6	0.34	0.29	Not Applicable

Interpreting the results, we can conclude that coefficients for HDD, CDD, and intercept parameter are significants (p-value less than 0.001). We found slight significance for β_5 coefficient (p-value between 0.1 and 0.05. We don't find significance for PREC (precipitations) or $\frac{nominalGDP}{GDPdeflator}$ coefficients.

Taking into account nominal GDP and population

Other option is to take into account nominal GDP and some regressors related to the steady-state, such as the Spanish population growth rate, since the estimation period falls within the long term. The following equation describes this case.

$$\begin{aligned} ENERGY\_DEMAND_t = \beta_0 + \beta_1 HDD_t + \beta_2 CDD_t + \beta_3 PREC_t + \\ \beta_4 nominalGDP_t + \beta_5 log(\frac{SpanishPopulation_t}{SpanishPopulation_{t-1}}) \end{aligned} \tag{3.14}$$

However, this model also does not improve the initial results as shown below.

Table 3.9 MSE adding information about nominal GDP and population growth rate

Base Temp	Linear MSE test	SVR MSE test	Deep Learning MSE test
14.5	1,091,721	166,109,399	940,805
15.5	1,088,966	7,652,457	1,036,323
16.6	1,086,226	432,212,828	1,289,250

Table 3.10 R^2 adding information about nominal GDP and population growth rate

Base Temp	Linear R^2 test	SVR R^2 test	Deep Learning R^2 test
14.5	0.37	Not Applicable	0.46
15.5	0.37	Not Applicable	0.40
16.6	0.37	Not Applicable	0.26

Interpreting the results and using the linear model and the Neural Networks, we can conclude that coefficients for HDD, CDD, and intercept parameter are significants (p-value less than 0.001). SVR obtain the worse results with this model. We do not find additional relations.

3.4 Simulating energy demand for 2020

We simulated a scenario that assumes that the IPI in Spain will grow continuously in the future, repeating previous growth rates. The simulation process takes into account the following hypothetical scenario:

1. IPI: It is assumed that Spain began a process of economic activation resulting in an increase in industrial activity. Therefore from January 2016 we will begin repeating (backwards) the data we have from December 2011 to November 2008. This implies the assumption that, in December 2019, the value of IPI will be 127.39. Our assumption is based on the evolution of IPI over the last years. The average annual growth rate for the period January 2014 to November 2016 (last data available) has been 2.24. The average annual growth rates for 2015 and 2016 have been3.24 and 2.04, respectively. These values lead us to trust our assumption.

2. Climatological data: We have assumed that weather will not change over the previous two years. Therefore, for each month and for the HDD, CDD and rainfall variables we take means of the corresponding month in the last two years. CDD and HDD have been calculated with a base temperature value of 15.5 degrees Celsius.

We are not assuming the direction of the time series relationship. We are just assuming a counterfactual and we try to evaluate our assumptions using this counterfactual.

Always according to the proposed scenario, the Deep Learning Model estimates that energy demand in December 2019 will be 23,109.17 GWh. If we use a base temperature of 14.5 degrees Celsius the prediction for energy demand in December 2019 is 22,818.62 GWh

while with a value of comfort of 16.5 for the base temperature, the expected energy demand according to the model's prediction will be 24,364.65 GWh.

The amount of CO_2 emitted is important for its environmental impact. Therefore, this paper also includes a small exercise about this issue. According to the Electricity Observatory data of World Wildlife Fund (WWF), in December 2015, Spain had issued an average of 0.269 kg of CO_2 per KWh consumed. Thus, if we consider an expected demand for December 2019 in the average scenario of 23,109.17 GWh, and assuming that the ratio of CO_2 emissions is maintained, Spain will emit into the atmosphere more than 6.2 million tons of CO_2. Introducing uncertainty in the scenarios based on the temperature of comfort, the interval of emissions will move from 6.1 to 6.5 million tons. Given that the Paris agreement on emissions will enter into force by the end of 2016, there is place for technology, for the energy mix and for efficiency to achieve our commitments.

3.5 Conclusions

We can observe the following facts in the demand for energy in Spain in recent years:

1. Since March 2007, according to available data, the monthly energy demand has been higher than the amount we simulated for December 2019 only 11 times. But since January 2013, only in July 2015 energy demand exceeded this level, reaching a value of 23,476 GWh.

2. According to a report by REE, in July 2015, generated of electricity from renewable energy sources reached 30.7% of the total energy produced. However, this figure includes renewable thermal energy (1.8%), which is obtained by burning waste and is thus a contaminant. In any case, there is a potential to reach the goal of 30% energy produced from renewable sources.

In 2005 electricity demand in the Iberian Peninsula amounted to 246,187 GWh. While electrical consumption in the islands was 14,517 GWh amounting to a total of 260,704 GWh. for the whole of Spain. On the other hand, we believe that the ratio of CO_2 emissions per KWh consumed in 2005 is higher than the ratio expected for 2019. This is because electricity technologies continue evolving towards more sustainable production methods, for example, increasing renewable energy sources. If the emissions in 2019 is 0.269 kg of CO_2 per KWh consumed) we predict that the volume of CO_2 emissions at the end of 2019 will correspond to around 75 million tons (between 6.1 and 6.5 million tons per month).

We think that this kind of exercise evidence illustrates the need for detailed, downloadable, easily usable and validated open data about energy consumption and emission that allows

us to analyze whether the initial objectives about greenhouse gas emissions are going to be achieved. As a preliminary conclusion, our evidence suggests that Spain may be on track to meet its commitments to European Union. These agreements imply that, by 2020 between 20% and 30% of the consumed energy comes from clean and renewable energy sources. We have serious doubts about how much could be achieved in reducing pollutant gases. However, the clean and renewable energy source development can be achieved by external factors to the regulation itself or, by developing energy policies aimed at the introduction of a sustainable mix (following the path initiated in the early 2000s), by the development of real sectors energy-related where Spain was considered highly innovative in terms of technology and by consumer awareness with improving energy efficiency. This productive model should boost activity and employment in relation to energy efficiency and clean energy production.

Chapter 4

Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy

4.1 Introduction

Public policy plays a critical role in regulating relationship between companies, investors and society. The importance of public policy for long-term investors has grown in recent years, due to [57]:

- Legislative reform of the financial sector in the wake of the global financial crisis.
- Governmental need for investors as a source of long-term growth.
- The increasing impact of environmental, social and governance factors on the ability of investors to deliver long-term returns.

So, the financial reform, the economic recovery and, in general terms, sustainability, are three factors that have been driven the public policy up the investor agenda in recent years. Sustainability involves considering issues like climate change, energy security, human rights, income inequality, international development and water stress, for instance. According to [57], policymakers seeking to work with long-term investors need to understand their needs and interests, encourage investors and other financial sector organizations to dialogue, facilitate long-term investor input to policy discussions, develop relationships with key long-term investment organizations and demonstrate commitment to long-term investor engagement. Investors (individuals or groups) are already engaged in public policy debates.

For example, in Europe, the Institutional Investors Group on Climate Change (IIGCC) brings "... investors together to use their significant collective influence to engage in dialogues with policymakers, investors and companies to accelerate the shift to a low carbon economy". Investment by these agents depend on markets that should be stable, well regulated, transparent and fair. The growing demand for affordable, reliable, domestically sourced, and low-carbon electricity is on the rise. It "is driven in part by evolving public policy priorities, especially reducing the health and environmental impacts of electricity service... Well-designed markets encourage economically efficient solutions, promote innovation and minimize unintended consequences" ([13]).

Policy objectives and new technologies are changing wholesale market design. A relevant case is the Spanish solar energy policy. A series of regulatory reforms since 2010 reduce revenues to existing renewable power generators and end the previous system of support to new renewable generation. This policy change has caused several claims by various organizations and altered the composition of the energy market. At the end, the Royal Decree of October 2015 strongly affected the solar energy market.

The analysis of the public opinion about specific government measures may be a useful component for long-term investment decisions. Generally, the main drive for investment decisions is the rate of return. But the public opinion can exert influence on politicians at several stages of their decision process. At the time of elections, for instance, candidates must respond to the public perception of their campaigns and it could affect the design of their political programs. Financial investors can also be influenced by the positive and negative perceptions about the company. In any case, if the government were interested in investments as a fundamental source of long-term growth, it could also be interested in analyzing the state of the public opinion. In this paper, we analyze the public opinion about energy policy of the Spanish Government using the Global Database of Events, Language, and Tone (GDELT). The GDELT Project ([27]) consists of over a quarter-billion event records in over 300 categories covering the entire world from 1979 to present. Our aim is to build sentiment indicators arising from this source of information and, in a final step, evaluate if positive and negative indexes have any effect on the evolution of key market variables as prices and demand. We do not try to evaluate whether there are causal effects from the sentiment indicators to the market variables but our purpose is to detect the existence of correlation among those variables.

The rest of the chapter is structured as follows: First, we provide a brief summary of the GDELT Project and explain the relation with the Big Data paradigm. In next section, we explain the tools, methods and techniques used in this study. Then, the results obtained

are discussed. The chapter ends up providing some policy implications and ideas for future research.

4.2 The GDELT Project

The GDELT Project, supported by Google Ideas, share real-time information and metadata with the world. This codified metadata (but not the text of the articles) is then released as an open data stream, updated every 15 minutes, providing a multilingual annotated index of the information. It includes broadcast, print, and online news sources. The project shares a database with trillions data points. Although, data is available as downloadable CSV files, few users have the storing capacity and processing power to download terabytes of data, and effectively query and analyze it. Google's BigQuery platform provides a way to interact with this huge information source. GDELT is a clear example of Big Data, while Google's BigQuery is an example of Infrastructure As a Service (IaaS) technology. According to [20], Big Data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of our database architectures. To gain value from this data, one must choose an alternative way to process it. Big Data technologies have huge variety of sources, huge volume of information – so much less time is needed to process information thanks to parallel processing and clustering infrastructure. GDELT maintains the GDELT Event Database, and the GDELT Global Knowledge Graph (GKG). The GKG begins April 1, 2013 and "... attempts to connect every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day" ([27]). The data files use Conflict and Mediation Event Observations (CAMEO) ([24]) coding for recording events. GKG also provides event identification (EventIDs) of each event found in the same article as the extracted information, allowing rich contextualization of events.

4.3 Methodology

In this work, we have used GKG table on Google's BigQuery platform, GKG table provides the "Themes" attribute, the list of all themes found in the document. We want to filter documents related to, at least, one of these two themes: "ENV_SOLAR" (which refers to solar power in general), and "FUELPRICES" (which refers to cost of fuel, energy and heating). The theme attribute is not available for the Event table. At the same time, we have

looked for events that refer to Spain at some point using the "Locations" attribute (which contains a list of all locations found in the text). In summary, we are using GKG table to filter information about solar power or cost of fuel, energy and heating and related (in some way) with Spain. Attribute "V2Tone" allows us to analyze the average "tone" of the document as a whole. The score ranges from -100 (extremely negative) to +100 (extremely positive). Common values range between -10 and +10, with 0 indicating neutral. This is calculated as Positive Score minus Negative Score. Positive Score is the percentage of all words in the article that were found to have a positive emotional connotation. Negative Score is the percentage of all words in the article that were found to have a negative emotional connotation. Big Query allows interaction with the whole GDELT dataset using Structure Query Language (SQL). An account in Google Cloud Services and activate Google's Cloud Storage to export and download data is required.

R ([47]) has been used to analyze and process data. Downloaded data from Google's Cloud Storage have been imported into R. After that, we have done a basic and exploratory analysis of the downloaded data. The analysis shows that there are some references, documents or URL's that are not related to energy policy. For example, some entries refer to scientific news from Canary Institute of Astrophysics ("ENV_SOLAR" theme). For this reason and for a more efficient measurement, we feel that it is necessary to analyze the text of the news and look for references to Spanish Government, council, ministry or ministers. This is a computation intensive task because it requires to deal with HTML tags and extract the text of the document. In the next section, we detail this process and explain different alternatives to improve execution time. Next and for each theme, we have grouped by day all mentions and calculated the mean tone and typical deviation in tone per day. At this stage, we only have evaluated documents written in Spanish or English because we need to find mentions to Spanish Government (Spanish and English have been the chosen languages to process, other languages will be included in future versions). The results have been placed into context with the policies that the government of Spain has implemented. Finally, we have applied a Correlated Topics Models (CTM) algorithm ([6]), based on Latent Dirichlet Allocation (LDA) algorithm ([7]), on the words contained in the mentions, news or documents written in Spanish. The rest of the dataset has not been included in this part of the study. We have not mixed languages, words with similar meaning but from different languages can be placed on different topics because writing is different. We leave the evaluation of other languages for future research. The "topicmodels" R package ([28]) allows us to execute LDA and CTM algorithms.

LDA allows to discover topics in large data collections described via topics. It does not require labeled data (unsupervised learning) and uses a stochastic procedure to generate the

topic weight vector. LDA represents documents as mixtures of topics that spit out words with certain probabilities. It is a bag-of-words model. For this reason, LDA can be used for document modelling and classification. LDA fails to directly model correlation between the occurrence of topic and, sometimes, the presence of one topic may be correlated with the presence of another (for example: "economic" and "business"). CTM is very similar to LDA except that topic proportions are drawn from a logistic normal distribution rather than a Dirichlet distribution. Applying the "topicmodels" R package to our dataset, we can obtain a list of words for every topic and, also, check the correlation between the topics obtained.

4.3.1 Getting the text of the documents

As we have mentioned before, getting the text of the documents is a compute-intensive phase because it requires to deal with HTML tags and extract the text of the document. This task is done in the following steps:

1. First, documents are accessed through its URL.

2. We detect the text language using "textcat" R package ([36]). If the document is written in Spanish or English we download the text and confirm that the text contains one of the following words: "gobierno", "government", "council", "ministers", "ministry", "ministro", "ministerio". In other case, we consider that the text does not mention Spanish Government. One should note that Spain location is referenced in the text according to GKG metadata.

3. For all downloaded texts, we clean the text removing HTML tags and stop-words (Spanish and English) in order to improve accuracy and performance. This task reduces the text size. Stop-words refer to the most common words in a language but given our aims they do not add value to the analysis of the topic.

Sequential and parallel execution modes have been tested here. A parallel algorithm, as opposed to a traditional sequential algorithm, is one which can be executed a piece at a time on many different processing p devices or processors, and then put back together again at the end to get the correct result.

The usefulness of this type of parallelization is that once the program structure is known, a few changes must be made to the program to be executed by several processors, and not as a distributed algorithm in which, first, we must establish the optimal communication structure. This usually involves making substantial changes to the program.

Two parallelization schemes have been evaluated. In the first one, we take advantage of multiple cores in one computer. The second parallelization method employs a master-slave

architecture ([9]). This architecture features a single processor running the main algorithm (master) which delegates the mission of getting the text among a group of processors (slaves). Slaves are responsible for processing URLs and getting the text and communicating results to the central process. In any case, if we have p processors, the original dataset is divided into p chunks, one processor processes only one of these chunks. In all cases, we have used a computer with Pentium V quad core and 8 GB RAM, managed by Operating System Centos 6.4. The Internet broadband speed is, roughly, 20 Mbps (download speed). Performance are usually measured in terms Speedup (S_p) and Efficiency (E_p):

$$S_p = \frac{T_1}{T_p} \tag{4.1}$$

$$E_p = \frac{S_p}{p} \tag{4.2}$$

where p is the number of processors, T_1 is the execution time of sequential algorithm and T_p is the execution time of the parallel implementation on p processors. The Simple Network of Workstations (*snow*) package ([60]) allows executing parallel code in R. It requires loading the code, loading the snow library, create a snow cluster (or execute in local mode using multicore CPU) and running the code, maintaining this order. Snow library can be used to start new R processes (workers) in our machine. The snow package is a scatter/gather paradigm, which works as follows:

1. The manager partitions the data into chunks and parcels them out to the workers (scatter phase).

2. The workers process their chunks.

3. The manager collects the results from the workers (gather phase) and combines them as appropriate to the application.

Snow can be used with socket connections, Message Passing Interface (MPI), Parallel Virtual Machine (PVM), or NetWorkSpaces ([60], [41]). The socket transport does not require any additional packages, and is very portable. We have used socket connections. Snow is a non-shared-memory system example, if we are using a network of workstations, each workstation has its own and independent memory. But, in the multicore and one-computer case, the memory is shared between all the running processes. In both cases, the cost of communications should be kept in mind. The cost of communication is dependent on a variety of features including the programming model semantics, the network topology, data handling and routing, and associated software protocols. Reducing the computation time by

adding more processors would only improve marginally the overall execution time as the communication costs remains fixed

4.4 Results

In this section we are going to summarize the main results of the different experiments. First, we show results from GDELT GKG.

4.4.1 Results using data from GDELT GKG

We are analyzing tone in mentions from GKG database. All mentions have some common characteristic, they refer to Spain as location at some point, themes "ENV_SOLAR" or "FUELPRICES" are detected in the text and, the lines contain one of the following words: "gobierno", "government", "council", "ministers", "ministry", "ministro", "ministerio" (we interpret it as the text mentions the Spanish Government). GDELT 2.0 and GKG new version are relativity recent. For this reason, we only have data filling these requisites from February 18th, 2015 to October 28th, 2015 (note that GDELT is a project in constant update, so we refer here data obtained in our interaction with Google's BigQuery on October 28th, 2015). The following figures show the mean and typical deviation in mentions (tone). All mentions (mentions without filtering words nor languages) and only government mentions are displayed and compared. Blue bars represent mean values in tone, black ones represent error bars. According to Figure 1 and its histogram in Figure 3, the average index of the mentions due to fuel and energy prices are negative indicating that the sentiment of news related to the solar energy policies is negative through this period. We must remember that the government introduced in October 2015 what was named as solar tax ("impuesto al sol") regulating consumption made by consumers who produce their own energy through photovoltaic systems. The discussions at the media did not began at the time of publishing the Royal Decree on October the 9th, 2015 but several months before as soon as the agents knew government's intention. It is not strange that the sentiment of the agents producing news is negative. On the other hand, when we include the word government as a control to build the sentiment index, the average as presented in Figures 4.2 and 4.4 is still more negative. So, the agents (producers, consumers, etc.) clearly express a negative reaction towards the fuel and energy prices and we associate it to the regulations in the energy market referred to these variables. The opinion expressed in other surrounding countries of Europe and also by the authorities of the UE was also negative towards the regulation.

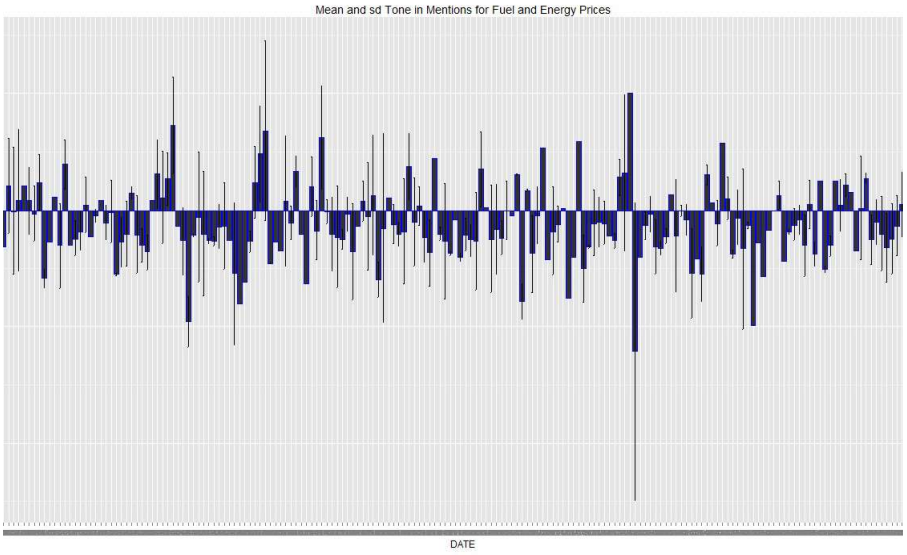

Fig. 4.1 All mentions for "FUELPRICES" theme in Spain

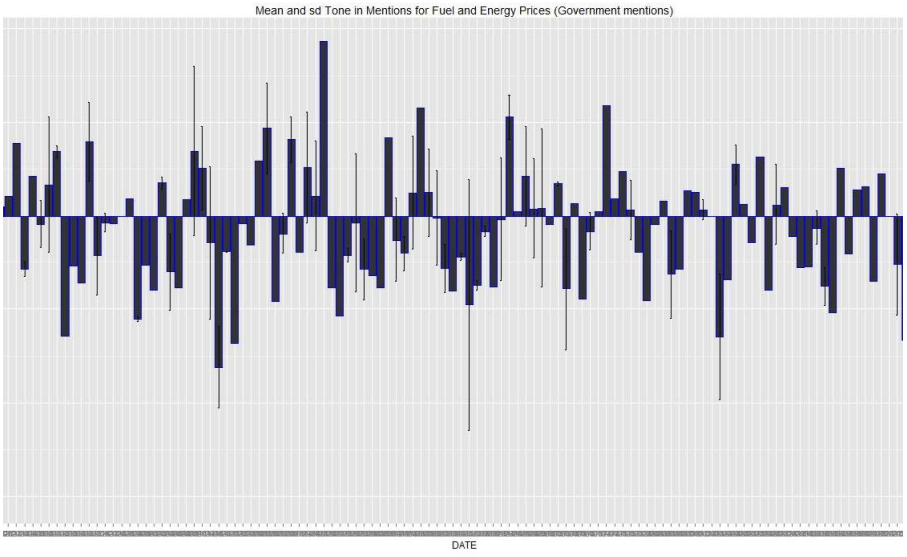

Fig. 4.2 Mentions for "FUELPRICES" theme in Spain filtering words (government mentions)

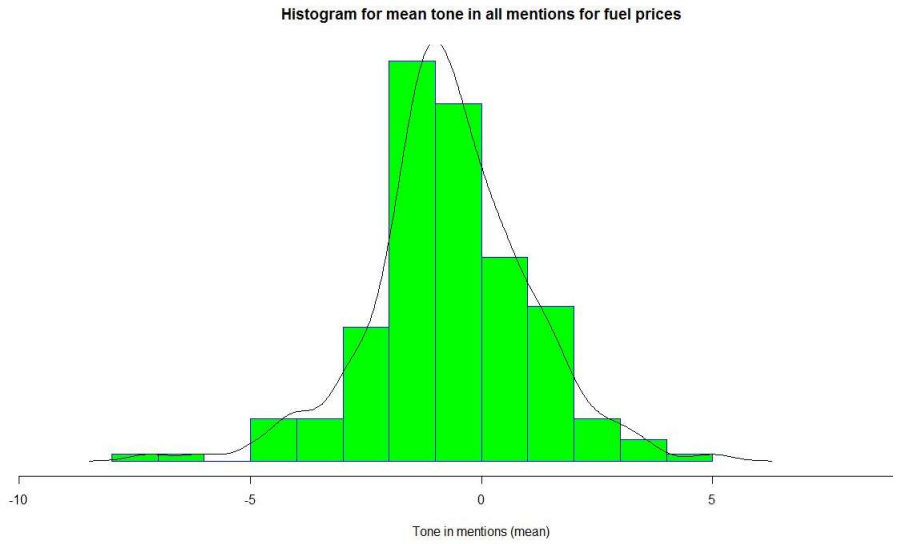

Fig. 4.3 Histogram - All mentions for "FUELPRICES" theme in Spain

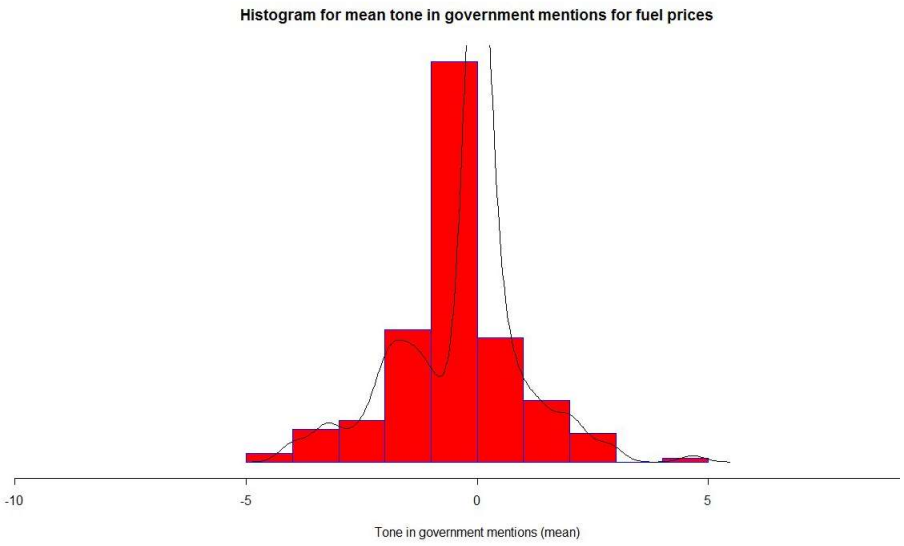

Fig. 4.4 Histogram - Mentions for "FUELPRICES" theme in Spain filtering words (government mentions)

In order to be able to use these data and conduct some test on them, we first check whether the indexes are normally distributed. Figures 4.5 and 4.6 present Q-Q plots, which are probability plots, i.e., a graphical method for comparing two probability distributions by

plotting their quantiles against each other. Here, we use Q-Q plot to compare data against Normal Distribution with mean and standard deviation according to the sample. Formally, the Shapiro-Wilk test ([52]) allows us to reject normality. For all data samples mean values are near to zero while typical deviation values are between 0.7 and 1.7. A normal distribution is symmetric about its mean, but this is not the case, and, taking into account the figures, we detect some extreme positive values which are balancing out a more frequent negative values and, for this reason, the mean tone in close to zero.

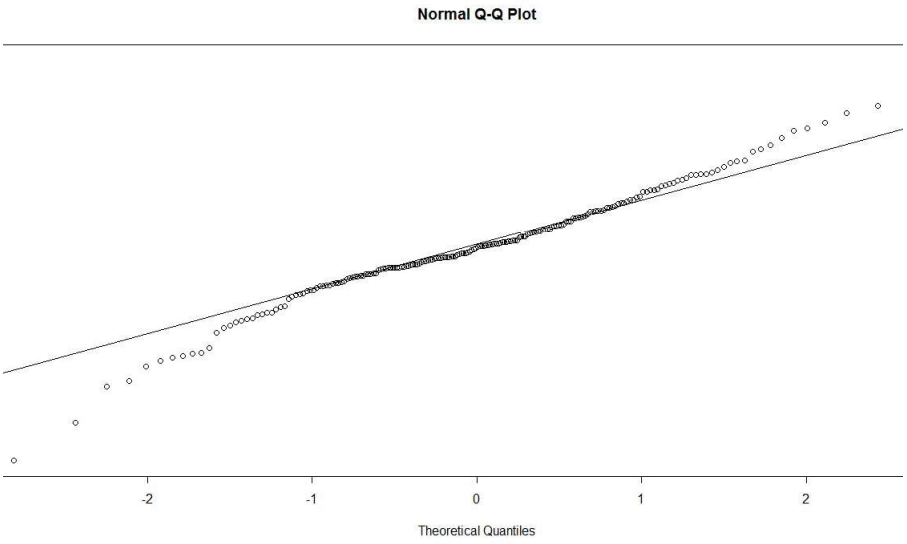

Fig. 4.5 Q-Q Plot - All mentions for "FUELPRICES" theme in Spain

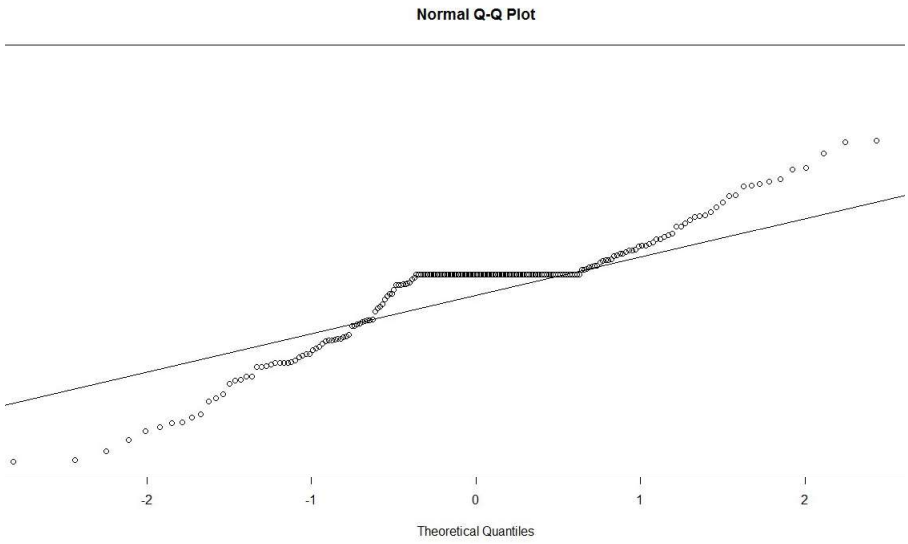

Fig. 4.6 Q-Q Plot - Mentions for "FUELPRICES" theme in Spain filtering words

Next, we conduct a similar exercise but filtering in GDELT a different theme than before. So, we include environment and solar ("ENV_SOLAR" theme) to the previous exercises and analyze tone in the same way. We can see than the sentiment index does provide some negative tone messages when we do not filter using words related to the government. However, once we filter for words related to the government, the negative tone appears much more clearly.

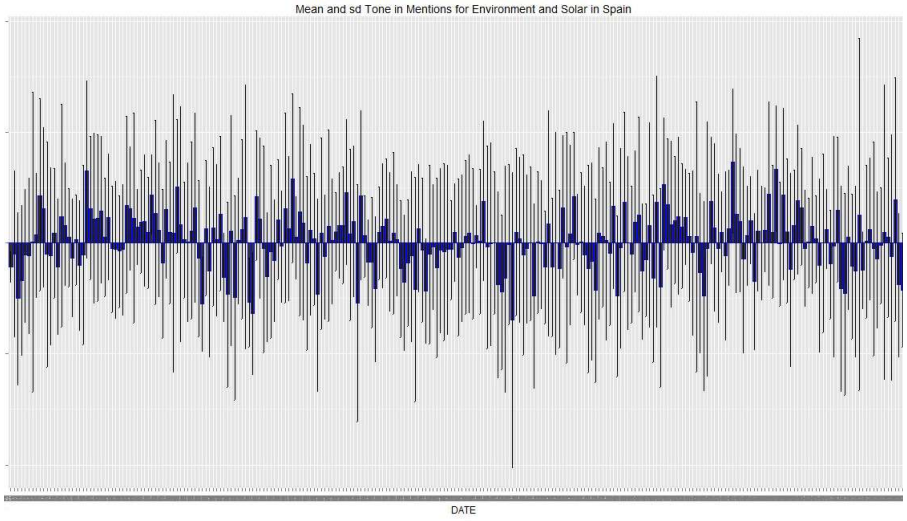

Fig. 4.7 All mentions for "ENV_SOLAR" theme in Spain

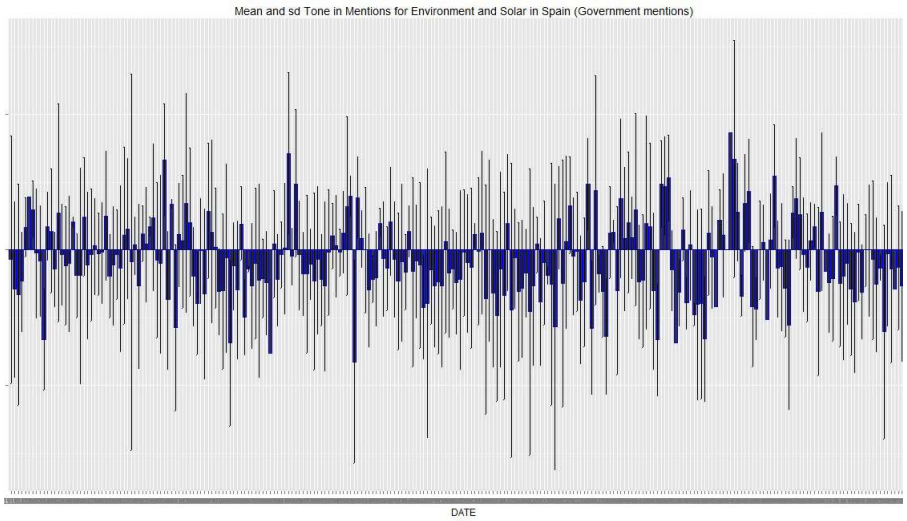

Fig. 4.8 Mentions for "ENV_SOLAR" theme in Spain filtering words (government mentions)

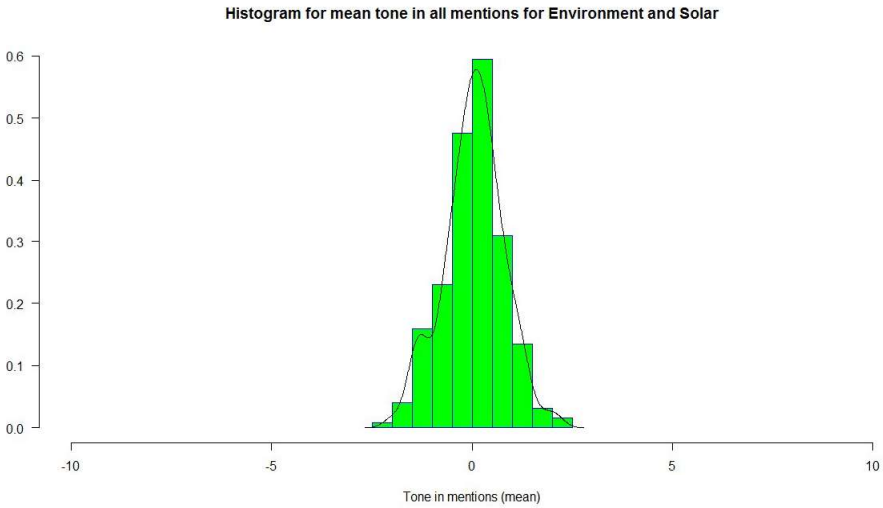

Fig. 4.9 Histogram – All mentions for "ENV_SOLAR" theme in Spain

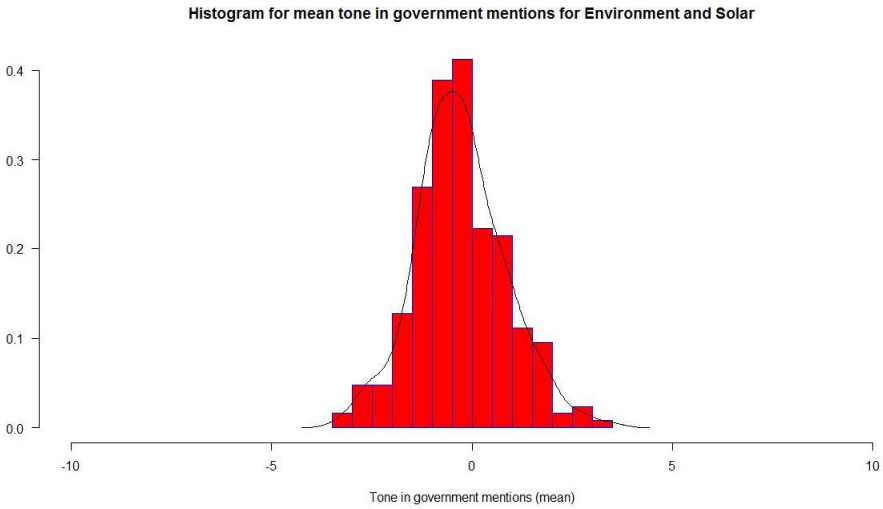

Fig. 4.10 Histogram – Mentions for "ENV_SOLAR" theme in Spain filtering words (government mentions)

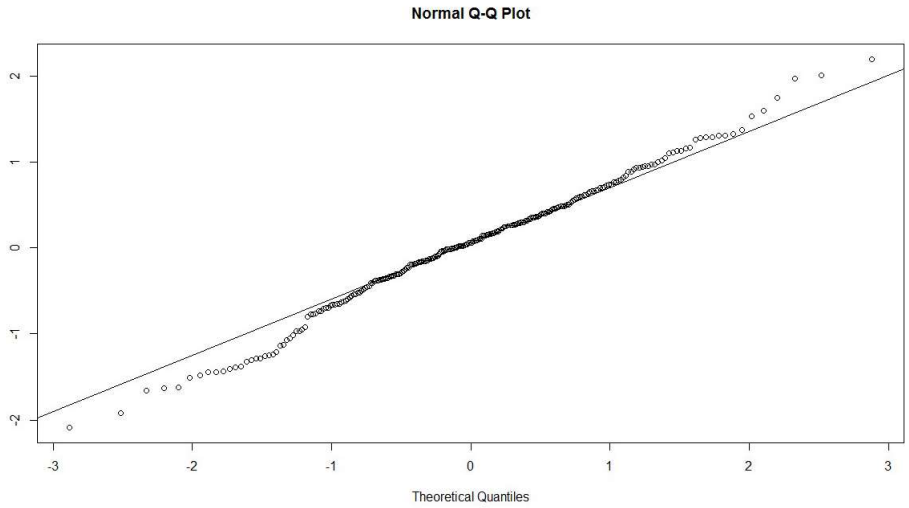

Fig. 4.11 Q-Q Plot – All mentions for "ENV_SOLAR" theme in Spain

Despite the graphic, Shapiro-Wilk normality test produces no suspicion about normality. But the mean tone, although is close to zero, is negative.

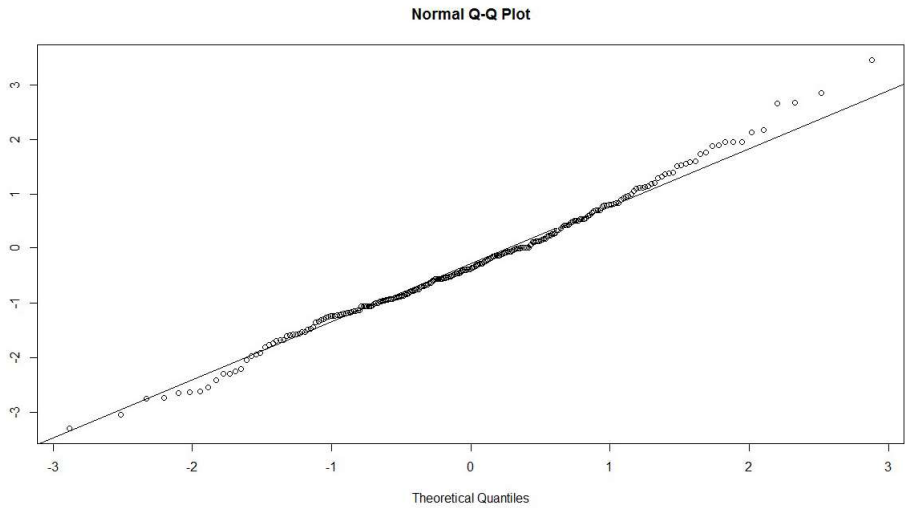

Fig. 4.12 Q-Q Plot – Mentions for "ENV_SOLAR" theme in Spain filtering words

4.4.2 Correlation analysis: Prices and Demand

We can obtain historical data about electricity prices and demand from OMIE ([46]). OMIE manages the electrical market for Spain and Portugal (MIBEL Market). We have used data from February 18th, 2015 to October 28th, 2015 (similar to GDELT data). Our aim is to evaluate whether there is any type of correlation between prices or demand in the MIBEL market and the mean tone of public opinion evaluated thanks to GKG data. For prices and demand, we have taken natural logs. Variables in Figure 4.13 must be interpreted in the following way: MeanALLSolar (ENV_SOLAR theme for Spain) does not include references to Spanish Government, MeanGovSolar does include it. Interpretation is similar for MeanALLFuel and MeanGovFuel (for FUELPRICES theme in this case). LogPrices and LogDemand refer to logarithm mean and daily values for prices and demand (respectively) from MIBEL historical data and the same period.

	MeanALLSolar	MeanGovSolar	MeanALLFuel	MeanGovFuel	LogPrices	LogDemand
MeanALLSolar	1.000000000	0.413180282	0.017009909	-0.016133573	-0.03178318	0.004742933
MeanGovSolar	0.413180282	1.000000000	-0.002291875	-0.044197743	-0.04209694	0.030167916
MeanALLFuel	0.017009909	-0.002291875	1.000000000	0.422778000	-0.07082776	-0.044638906
MeanGovFuel	-0.016133573	-0.044197743	0.422778000	1.000000000	-0.01492889	-0.006255875
LogPrices	-0.031783179	-0.042096940	-0.070827763	-0.014928895	1.00000000	0.423427846
LogDemand	0.004742933	0.030167916	-0.044638906	-0.006255875	0.42342785	1.000000000

Fig. 4.13 Correlations results

There is no robust evidence about correlation but we can observe that LogPrices and mean tone collected by MeanALLFuel and MeanGovFuel present evidence of some weak correlation. So, public opinion could to some extent weakly affect fuel prices and, indirectly, fuel demand in the short-term.

4.4.3 CTM results using CTM algorithm

In this subsection we present the results obtained using a CTM algorithm to discover and correlate topics. We must note that we only have applied the algorithm to Spanish texts. The R package "topicmodels" allow us to display the graphs collected in Figure 4.14. For the theme named "FUELPRICES" and text written in Spanish, the cluster Group 1 corresponds to HTML tags or other words that have not been properly removed or English words that appear in texts that "textcat" R package has been classified as written in Spanish. On the other hand, clusters named Group 3 and 6 refer to words like (translated from Spanish) "stock exchange", "market", "government", "income", "Europe", "state", "congress"..., "gas", "price", "growth" and other Spanish locations as "Madrid" or "Barcelona" are words contained in the rest of the groups.

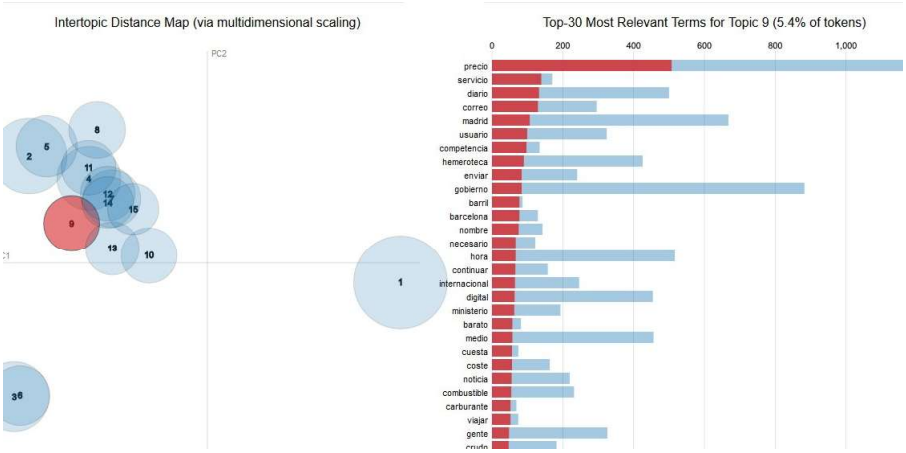

Fig. 4.14 Using "topicmodels" R package. "FUELPRICES" theme, text written in Spanish

For "ENV_SOLAR" and text written in Spanish, Group 1 is similar to the last case. Group 5, 12 and 15 refer to words like "solar", "system", "electricity", "law", "change", "tax", and months (written in Spanish).

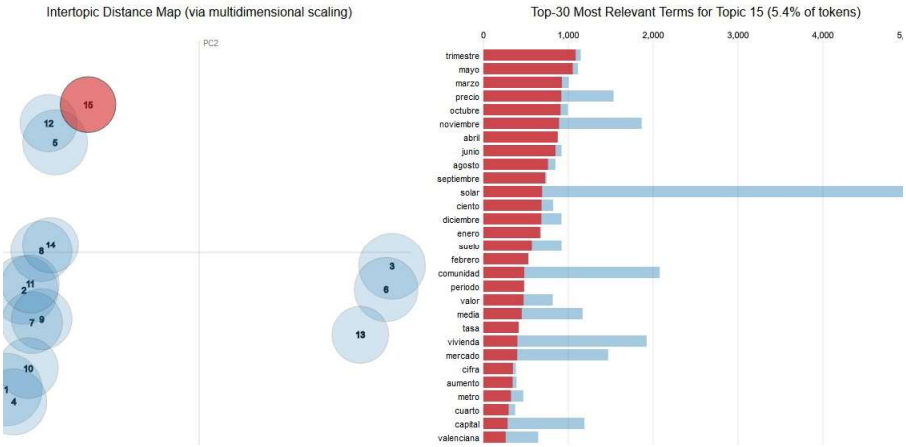

Fig. 4.15 Using "topicmodels" R package. "ENV_SOLAR" theme, text written in Spanish

The CTM model allows to classify documents or news in media. We also think that we will be able to use it in the future to do further correlation analysis. Of course our final aim will be to use this technique in future research to do causal analysis from the information collected (the indexes built on it) and the movement of key variables in energy markets.

4.4.4 Speedup and Efficiency analysis (getting the text)

As we have explained in the methodology section, getting the text from URLs is a compute-intensive phase. We present two figures for analyzing sequential and parallel execution modes. They summarize the performance in terms of Speedup and Efficiency according to equations (4.1) and (4.2). Figure 4.16 shows that speedup improves when using a network of workstations. Although efficiency (in terms of reducing execution time) increases with multicore execution a network of workstations is still preferred:

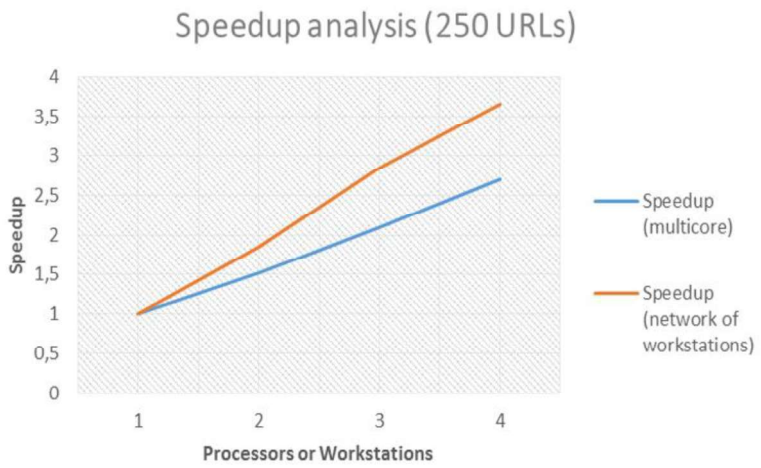

Fig. 4.16 Speedup comparison

Our code does not require the dispatch of regular data between processes. Therefore, when we are using a network of workstations, the communication cost is not excessive and efficiency can be maintained at a constant level. However, in the multicore case the computer memory has to be shared and this issue impacts in the efficiency values.

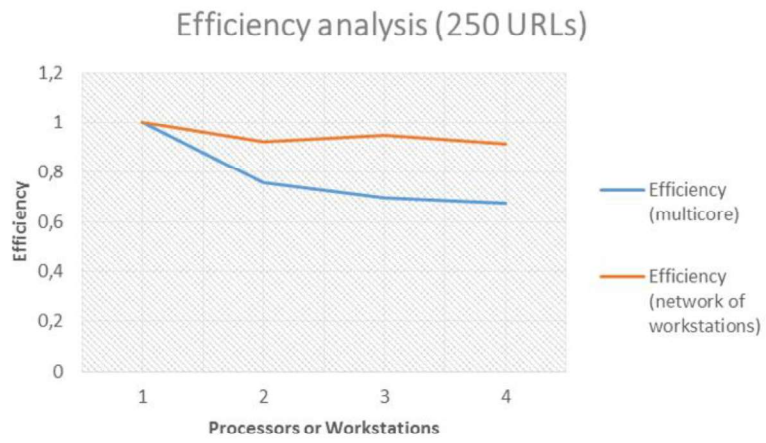

Fig. 4.17 Efficiency analysis

A multicore execution can be used to reduce execution time. Nevertheless, a network of workstations is preferred.

4.5 Conclusions and future work

In this chapter we have used extensive data from several sources to analyze two issues related to energy markets. First, we analyze the public opinion about energy policy of the Spanish government using GDELT. Second, we conduct a correlation analysis between sentiment variables about the public policy and real prices and demand taken from the MIBEL energy market for the same period. Two results are worth emphasizing. On one hand, we detect negative feelings about the solar energy policy introduced by the Spanish government in 2015. On the other, hand, we find weak correlation between the indexes (tone) in mentions from GKG database and average daily log prices of energy. We do not find any correlation to average daily energy demand. There are many extensions using extensive databases like the one in this paper or similar to follow different research lines in the future. We only quote two possibilities closely related to this exercise. First, we have only taken into account Spanish or English while alternative languages could be important to build sentiment indexes. Second, we have only presented correlation analysis between the indexes and average prices and demand but some formal demand model where to include these indexes as explanatory variables is necessary to accurately measure the potential of these variables to explain the evolution of key energy market variables.

Chapter 5

Conclusiones

5.1 Conclusiones y trabajo futuro

Los objetivos que se han planteado en esta tesis han sido los siguientes:

1. Analizar la relación entre precio y demanda de energía eléctrica en el mercado energético de la península Ibérica. Se pretende concluir cuáles son las variables que más influyen en los incrementos o decrementos del precio.

2. Proponer mecanismos que permitan evaluar si España será capaz de cumplir con las obligaciones incurridas con la EU en lo que se refiere a la reducción de gases de efecto invernadero.

3. Analizar desde una perspectiva técnica la opinión que la comunidad pública ha manifestado sobre ciertas actuaciones del gobierno español en materia energética, contribuyendo a la difusión de técnicas que pueden ayudar a una mejor evaluación de la opinión que el mercado tiene sobre las acciones del ejecutivo.

El estudio realizado ha concluido que un incremento en precios afecta negativamente a un incremento de la demanda. Sin embargo, la relación detectada es muy débil y poco significativa. Añadir nuevas variables al estudio y una mayor profundidad de las series históricas sería necesario para validar los resultados. Este tipo de estudios podría claramente impulsarse mediante iniciativas abiertas, estandarizadas y coordinadas de datos abiertos. La creación de repositorios masivos de información estandarizada, que incluya elementos automáticos de consulta, impulsaría una política de transparencia y mejora al proporcionar medios para la evaluación continua.

También, se han generado distintas simulaciones que permite estimar la demanda energética en España en el año 2020. De esta forma y considerando que la relación entre

consumo de energía eléctrica y emisión de gases de efectos invernaderos se mantendrá a los niveles actuales, se ha analizado si España sería capaz de cumplir con los compromisos adquiridos con la Unión Europea en lo que se refiere a la reducción de gases contaminantes. Estos acuerdos imponen que entre el 20% y 30% de la energía consumida deben proceder de energías limpias y renovables. Los resultados muestran que el país se encuentra en condiciones de cumplir con sus compromisos.

Por último y empleando el repositorio de datos masivos GDELT, se ha construido un indicador de opinión para evaluar la opinión pública acerca de la reforma energética introducidas por el gobierno español y conocidas coloquialmente como "impuesto al sol". Se ha buscado también relacionar este índice con la evolución de los precios y la demanda. Mientras sí ha sido posible evidenciar un impacto negativo de la reforma en la opinión pública, no se ha podido demostrar que este impacto haya afectado de forma relevante a la demanda energética o precios asociados.

Para el estudio se ha empleado una metodología que introduce un amplio abanico de datos, pruebas y medidas de cara a justificar la fiabilidad y rigorosidad de los resultados obtenidos. El trabajo se ha enriquecido a través de empleo de técnicas muy diversas, dando prioridad a los modelos fácilmente interpretables. En la presente tesis han convivido modelos lineales clásicos, con algoritmos más frecuentes en la disciplina de aprendizaje automático. Ejemplos de estos últimos modelos son SVM, Random Forest, Algoritmos Evolutivos o Redes Neuronales. La metodología empleada se considera también una contribución notable de la tesis puesto que propone un conjunto de herramientas rigurosas y científicamente validadas que pueden ayudar a la evaluación de las políticas públicas en particular y análisis de la situación económica en general.

En un futuro, se propone enriquecer el trabajo realizado en base a las siguientes acciones:

- Añadir fuentes de datos adicionales que enriquezcan el modelo.
- Incorporación y evaluación de nuevos modelos para complementar el análisis. Por ejemplo, los Modelos Aditivos Generalizados (GAM) podrían resultar útiles en este contexto.
- Profundizar sobre la relación entre el mercado energético y las distintas industrias y sectores productivos. Pretendemos así llegar a resultados más concretos y desglosados de los presentados en este trabajo.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[2] Alfred J. Lotka (1925). Elements of Physical Biology. *Williams and Wilkins Company*, page 435.

[3] Arbib, M., Ballard, D., Bower, J., and Orban, G. (1994). *Neural Networks Algorithms , Applications*, volume 7.

[4] Arel, I., Rose, D. C., and Karnowski, T. P. (2010). Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *IEEE computational intelligence magazine*, 5(4):13–18.

[5] Bergstra JAMESBERGSTRA, J. and Yoshua Bengio YOSHUABENGIO, U. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305.

[6] Blei, D. M., Lafferty, J. D., Lafferty, D. M. B., and D., J. (2006). Correlated Topic Models. *Advances in Neural Information Processing Systems 18*, pages 147–154.

[7] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

[8] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

[9] Cantu-Paz, E. (1998). Designing Efficient Master-Slave Parallel Genetic Algorithms. In *Genetic Programming 1998: Proceedings of the Third Annual Conference*, number 97004, page 455.

[10] Capros, P., Mantzos, L., Parousos, L., Tasios, N., Klaassen, G., and Van Ierland, T. (2011). Analysis of the EU policy package on climate change and renewables. *Energy Policy*, 39(3):1476–1485.

[11] Chapelle, O. and Vapnik, V. (1999). Model Selection for Support Vector Machines. *Advances in Neural Information Processing Systems*, 12:??—-??

[12] Chapelle, O. and Vapnik, V. (2000). Model Selection for Support Vector Machines. In *Advances in Neural Information Processing Systems*, pages 230–236.

[13] Cochran, J., Miller, M., Milligan, M., Ela, E., Arent, D., and Bloom, A. (2013). Market Evolution: Wholesale Electricity Market Design for 21 st Century Power Systems. *Contract*, (October 2013):1–57.

[14] Coello, C. a. C., Lamont, G. B., and Veldhuizen, D. a. V. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems Second Edition*.

[15] Commission, F. E. R. et al. (2015). Energy primer, a handbook of energy market basics. *Federal Energy Regulatory Commission: Washington, DC, USA*.

[16] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[17] Dahl, C. (2009). Energy demand and supply elasticities. *Energy Policy*, page 72.

[18] Deaton, A. and Muellbauer, J. (1980). *Economics and consumer behavior*. Cambridge university press.

[19] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

[20] Dumbill, E. (2012). *Planning for Big Data*.

[21] fiware (2016). Fiware lab data portal statistics. [Accessed: 2018-02-25].

[22] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22.

[23] Fritsch, S., Guenther, F., and Suling, M. (2012). R packages - neuralnet. *Cran*, page 13.

[24] Gerner, D. J., Abu-Jabr, R., Schrodt, P. A., and Yilmaz, Ö. (2002). Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions (2002). Technical report.

[25] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, volume Addison-We.

[26] Goldberg, D. E. and Holland, J. H. (1988). Genetic Algorithms and Machine Learning. *Machine Learning*, 3(2):95–99.

[27] GOOGLE (2016). The gdelt project: Watching our world unfold. [Accessed: 2018-02-25].

[28] Grün, B. and Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13):1–30.

[29] Harris, C. (2013). *Electricity Markets: Pricing, Structures and Economics*.

[30] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. *Springer 2001*, 18(4):746.

[31] Hayashi, F. (2000). Econometrics. 2000. *Princeton University Press. Section*, 1:60–69.

[32] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

[33] Horton, N. J. and Kleinman, K. (2010). Linear regression and ANOVA. *Using R for Data Management, Statistical Analysis, and Graphics*, pages 137–171.

[34] James, G., Witten, D., Hastie, T., and Tibishirani, R. (2013). *An Introduction to Statistical Learning*.

[35] Joachims, T. (2006). Training linear SVMs in linear time. *Kdd*, page 217.

[36] Kurt Hornik and Patrick Mair and Johannes Rauch and Wilhelm Geiger and Christian Buchta and Ingo Feinerer (2013). The textcat Package for n-Gram Based Text Categorization in R. *Journal of Statistical Software*, 52(6):1–17.

[37] Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (1996). *Applied Linear Statistical Models*, volume Fifth.

[38] Labandeira, X., Labeaga, J. M., and López-Otero, X. (2017). A meta-analysis on the price elasticity of energy demand. *Energy Policy*, 102:549–568.

[39] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

[40] López, C. B., Carreras, A., and Tafunell, X. (2005). *Estadísticas históricas de España: siglos XIX-XX*, volume 3. Fundacion BBVA.

[41] Matloff, N. (2015). *Parallel computing for data science: with examples in R, C++ and CUDA*, volume 28. CRC Press.

[42] Meyer, D., Hornik, K., Weingessel, A., Leisch, F., and Davidmeyerr-projectorg, M. D. M. (2015). Package 'e1071'.

[43] Montana, D. J. and Davis, L. (1989). Training Feedforward Neural Networks Using Genetic Algorithms. *Proceedings of the 11th International Joint Conference on Artificial intelligence - Volume 1*, 89:762–767.

[44] NCC (2018). National centers for environmental information. national oceanic and atmospheric administration. [Accessed: 2018-02-25].

[45] Odum, H. T. and Odum, E. C. (2006). The prosperous way down. *Energy*, 31(1 SPEC. ISS.):21–32.

[46] OMIE (2016). Omi-polo español s.a. (omie): Market results. [Accessed: 2018-02-25].

[47] R Core Team (2014). R: A Language and Environment for Statistical Computing.

[48] REE (2016). Red eléctrica de españa. spain electrical energy report (in spanish). [Accessed: 2018-02-25].

[49] Ristinen, R. A. and Kraushaar, J. J. (1998). Energy and the environment. *Energy and the Environment, by Robert A. Ristinen, Jack J. Kraushaar, pp. 384. ISBN 0-471-17248-0. Wiley-VCH*, page 384.

[50] Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pages 1651–1686.

[51] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

[52] Shapiro, S. S. and Wilk, M. B. (1965). An Analysis of Variance Test for Normailty (Complete Samples). *Biometrika*, 52(3-4):591–611.

[53] Smola, A. J. A. A. J., Schölkopf, B., Sch, B., and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

[54] Srinivas, N. and Deb, K. (1994). Muiltiobjective Optimization Using Nondominated Sorting in Genetic Algorithms. *Evolutionary Computation*, 2(3):221–248.

[55] Srivastava, V. K. and Dwivedi, T. D. (1979). Estimation of seemingly unrelated regression equations. A brief survey. *Journal of Econometrics*, 10(1):15–32.

[56] Stigler, S. M. (1981). Gauss and the Invention of Least Squares. *The Annals of Statistics*, 9(3):465–474.

[57] Sullivan, R., Martindale, W., Robins, N., and Winch, H. (2014). *Policy Frameworks for Long-Term Responsible Investment: The Case for Investor Engagement in Public Policy*. Principles for Responsible Investment.

[58] Tiana, M. et al. (2012). El impacto de la crisis económica sobre la industria española. *Boletín Económico*, (NOV).

[59] Tibshirani, R. and Society, R. S. (1996). Regression and shrinkage via the Lasso. *J R Stat Soc, Ser B*, 58(1):267–288.

[60] Tierney, L., Rossini, A., Li, N., and Sevcikova, H. (2013). snow: simple network of workstations. r package version 0.3-13. *Retrieved*, 13:2015.

[61] Trautmann, H., Steuer, D., and Mersmann, O. (2010). mco: Multicriteria optimization algorithms and related functions. r package version 1.0. 9.

[62] Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231.

[63] Vandaele, W. (1981). Wald, likelihood ratio, and Lagrange multiplier tests as an F test. *Economics Letters*, 8(4):361–365.

[64] Wolpert, D. (1995). No free lunch theorems for search. *Most*, pages 1–38.

[65] Worrell, E., Bernstein, L., Roy, J., Price, L., and Harnisch, J. (2009). Industrial energy efficiency and climate change mitigation. *Energy Efficiency*, 2(2):109–123.

[66] Zanakis, S. H., Evans, J. R., and Vazacopoulos, A. a. (1989). Heuristic methods and applications: A categorized survey. *European Journal of Operational Research*, 43(1):88–110.

[67] Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57(298):348–368.

[68] Zhang, G., Eddy Patuwo, B., Y. Hu, M., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting*, 14(1):35–62.

[69] Zitzler, E., Deb, K., and Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms: empirical results. *Evolutionary computation*, 8(2):173–95.

Appendix A

Publications

This thesis has led to the following publications:

Bodas-Sagi, Diego J., and Labeaga, José M. (2018) "Predicting Energy Demand in Spain and Compliance with the Greenhouse Gas Emissions Agreements." In Modeling, Dynamics, Optimization and Bioeconomics III. Springer.

Bodas-Sagi, Diego J., and Labeaga, José M. (2016) "Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy." International Journal of Interactive Multimedia and Artificial Intelligence 3.6: pp. 38-43.