Modelling Energy Intensity: A Computational Social Science Approach

UC3M - Master in Computational Social Science

Eric Hausken-Brates

2024-06-21

Abstract

The most densely populated areas have the worst energy intensity. Wind energy generated has a positive relationship with energy intensity, meaning more wind energy is related to worse energy intensity. Only a couple months have a siginificant affect on predicting energy intensity. And different months do not have significant affect mean intensity overall. Only when combined with CCAA does it show an effect.

Table of contents

Introduction **4**

Literature Review **5**

Methodology **6**

Data Wrangling 6

Correlations 7

Linear Models 8

Mixed Models 9

Random Forest 9

Model #2 10

Regularized Regression 11

Results **13**

Is Month a Significant Predictor? 13

CCAA as a Significant Predictor 13

Are the survey results significant? 13

Discussion **16**

Interpretation of Results 16

Model Comparisons 16

Conclusion **18**

References **19**

Appendices 20

Appendix A: Data Preprocessing Steps 20

Appendix B: Detailed Model Outputs 20

Appendix C: Additional Figures and Tables 20

```r
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, fig.width = 7, fig.height = 4)
knitr::knit_hooks$set(inline = function(x) {format(x, digits = 4, big.mark = ",")})

library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(caret)
```

```
Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift
```

```r
library(pdp)
```

```
Attaching package: 'pdp'

The following object is masked from 'package:purrr':

    partial
```

```r
rm(list = ls())
load(".RData")

new_theme <- theme_classic()
theme_set(new_theme)
```

```r
theme_update(
  plot.background = element_rect(fill = "gray75"),
  plot.title.position = "plot",
)
```

Introduction

Energy intensity, defined as the amount of energy consumed per unit of economic output, is a crucial indicator of a region's energy efficiency. This thesis investigates the factors influencing energy intensity in various regions, focusing on the role of wind energy generation and population density. Understanding these relationships is essential for developing effective energy policies and improving sustainability.

2024-06-14

Literature Review

There have been many studies and papers about energy efficiency in Spain and in general. In one study from 2003, “Policy networks of wind energy: The story of the first commercial wind farm in Spain, the researchers investigated the development of the wind energy boom in Spain. They explained that one significant factor”has been the implementation of effective fiscal policies” (page 462). These policies “helped to lower the cost of electricity generated by wind power.” The paper goes on to describe the history of wind energy, starting in Andalucia in the 1980s, and the national government’s laws to liberalize the electricity market in the 1990s. The development of renewable energy in Spain was driven by the need to reduce the CO2 emissions and produce its own electricity instead of depending on oil-producing countries.

Several studies over the years have explored the mechanisms for energy efficiency. In “Directed Technical Change and Energy Intensity Dynamics: Structural Change vs. Energy Efficiency,” the authors decomposed the changes in energy intensity (energy consumption per GDP) into structural effect and efficiency effect. They found that an increase in energy price is related to a decrease in energy intensity. A variable that affects energy efficiency is technological change, which can help with producing electricity, but can also backfire when the efficiency causes more energy use. According to the paper, “Do Spanish efficiency actions trigger JEVON’S paradox?”, technological changes in energy efficiency possibly backfired in some sectors, meaning that energy intensity was not diminished despite the innovation. Other factors that can affect energy efficiently include improvements in processes rather than technological advancements.

Understanding Jevon’s paradox is critical for policy decisions surrounding renewable energy. As renewable energy sources like wind and solar become more efficient and cost-effective, it is crucial to monitor whether these improvements lead to increased overall energy consumption. This could occur if the reduced cost of energy from renewable sources stimulates higher demand or if the savings are redirected towards other energy-consuming activities.

Methodology

Data Wrangling

The dataset used in this study includes various variables such as monthly electricity generation and consumption, population, economic indicators, survey results, and climate data. The data was downloaded directly from official government websites and from using their APIs.

```r
wind_data_plot <- PIVOT_data_pop_pib_land_energy |>
  ggplot(aes(x = as.Date(datetime), y = percentage_Eólica)) +
  geom_line() +
  geom_smooth(color = "#000066") +
  facet_wrap(~CCAAnombre, scales = "free_y", ncol = 4) +
  scale_x_date(date_labels = "'%y", minor_breaks = NULL) +
  scale_y_continuous(labels = scales::label_percent(accuracy = NULL), minor_breaks = NULL) +
  labs(
    y = "Wind energy as percentage of total",
    x = NULL,
    #subtitle = "Total energy generated per CCAA, 2014-23, Monthly",
    #title = "No clear pattern for all CCAA, some increase, some decreasing",
    caption = "Data: Red Eléctrica"
  )

ggsave(filename = "wind-timeline.pdf", wind_data_plot,
       width =7, height = 7, units = "in")
```

```r
relationship_plot <- agg_data |>
    filter(date == as.Date("2022-12-01")) |>
  ggplot(aes(x = Eolica, y = energy_int, label = CCAAnombre)) +
  geom_text(
          vjust =  -0.5,
          hjust = ifelse(agg_data$Eolica[agg_data$date==as.Date("2023-11-01")] > 0, 0.5, 0),
          color = "#000066", size = 2,
          ) +
  geom_point() +
  scale_x_continuous(
    trans = "log10",
    labels = scales::label_number( scale_cut = scales::cut_long_scale())
  ) +
  scale_y_continuous(
  #  labels = scales::label_number(scale = 1)
  ) +
  labs(
    x = "Wind energy generated - MWh",
    y = "Energy intensity - GWh/Euro",
    caption = "Data sources: Red Eléctrica, INE"
  )

ggsave(filename = "relationship.pdf", relationship_plot,
       width = 7, height = 5, units = "in")
```

Correlations

Pearson's correlation coefficient was used to assess the relationships between variables. The correlation matrix (Figure 1) highlights the key relationships, indicating a log-linear relationship between wind energy and energy intensity.

```
cor_plot <- DataExplorer::plot_correlation(agg_data2022[,c(2,4,7,8,11,14:15,17,20:31,34:37)],
                                           geom_text_args = list(size = 2),
                                           ggtheme = theme_classic(),
                                           )
```

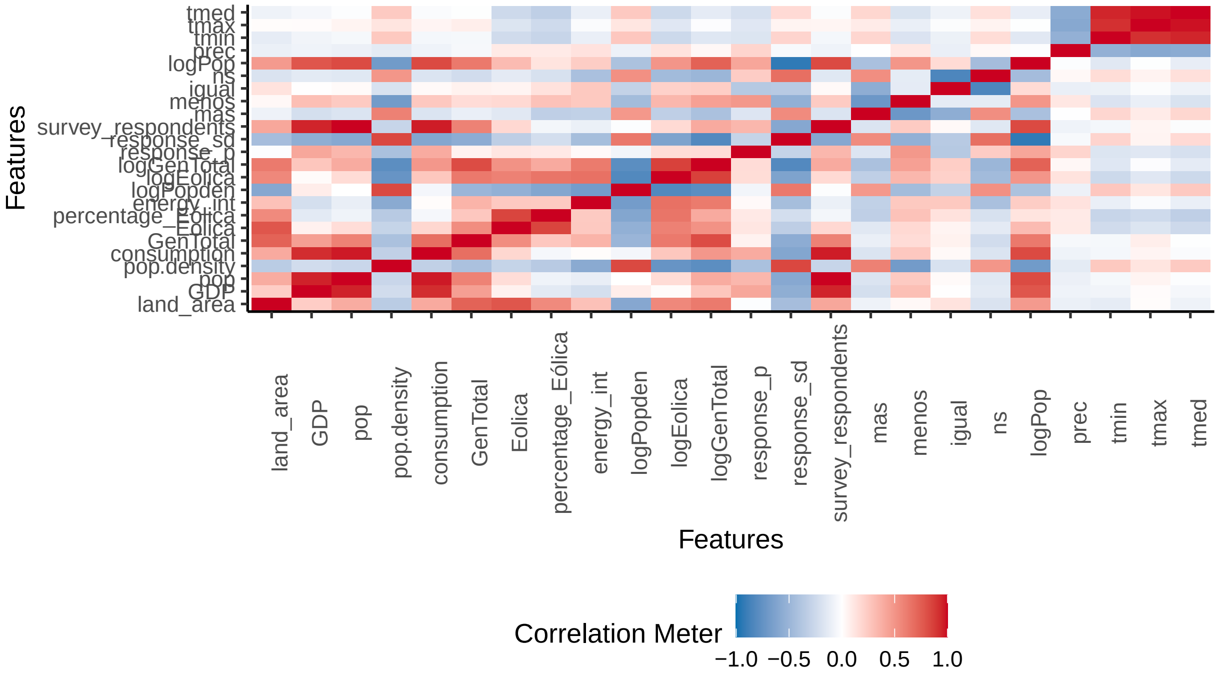


Table 1: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	216	5245.1719	NA	NA	NA	NA
2	204	2791.8502	12	2453.3217	143.20372	0
3	204	4803.2537	0	-2011.4036	NA	NA
4	198	282.6729	6	4520.5809	527.74489	0
5	209	480.0623	-11	-197.3894	12.56933	0

```r
ggsave(filename = "cor_plot.pdf", cor_plot, width = 9, height = 7, units = "in" )
```

Linear Models

Initial linear regression models were developed to explore the impact of wind energy and population density on energy intensity. The results suggested a significant positive relationship between log-transformed wind energy and energy intensity.

```r
month_plot <- agg_data2022 |>
  ggplot(aes(y = energy_int, x = logEolica, color = month)) +
  geom_line(alpha = .1) + geom_point(alpha = .1)+
  geom_smooth(method = "glm", se = T) +
  labs(
  #  title = "Mean energy intensity not significantly different across months"
  )

ggsave(filename = "month-plot.pdf", month_plot, width = 7, height = 4, units = "in")
```

```r
kableExtra::kbl(anova(aov_0, aov_1, aov_2, aov_3, aov_ccaa),
                format = "latex", row.names = T, caption = "Analysis of Variance Table",
        )
```

```r
for_tune_plot <- plot(for_tune, metric = "Rsquared")

coef(for_tune$finalModel, 3)
```

```
NULL
```

```r
coef(for_tune$finalModel, 6)
```

```
NULL
```

Mixed Models

Mixed-effects models were employed to account for the hierarchical structure of the data, with CCAA and month as random effects. These models provided a better fit compared to simple linear models.

```
lmer_basic
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: energy_int ~ logEolica + percentage_Eólica + (1 | CCAAnombre) +
    (1 | month)
   Data: agg_data2022
REML criterion at convergence: 849.3986
Random effects:
 Groups     Name        Std.Dev.
 CCAAnombre (Intercept) 5.316
 month      (Intercept) 0.915
 Residual               1.188
Number of obs: 228, groups:  CCAAnombre, 19; month, 12
Fixed Effects:
      (Intercept)          logEolica  percentage_Eólica
          17.9239            -0.2607            -1.0500
```

Random Forest

Random forest models were used to capture non-linear relationships and interactions between variables. The models indicated that certain regions, particularly Asturias, significantly influenced energy intensity.

I decided to run a *Random Forest* model without a variable for CCAA. This removal should show us which variables are important to predict energy intensity, without directly asking, "Which autonomous community is this?" Although one possibility would be only keep the binary dummy variable for when CCAA is equal to Asturias.

I also removed `month` because it has not shown to be important without CCAA involved.

This shrunken model with fewer explanatory variables had somewhat similar results. The most important variable was if the CCAA was Asturias. That was such a big outlier I had to keep it as a binary variable. The other most important variables were average temperature and the survey results for `igual`.

```
partial_ccaa <- partial(rf_tune, pred.var = "CCAAnombre", plot = F) |>
  ggplot(aes(x = yhat, y = CCAAnombre, xmin = 0, xmax = yhat)) +
  geom_linerange(color = "gray75", linetype = "dashed") +
  geom_point(color = "#000066") + # xlim(c(15,21)) +
  labs(y = NULL) + scale_x_continuous(oob = scales::squish, limits = c(14,21))
ggsave(partial_ccaa, filename = "partial_CCAA.pdf", width = 4, height = 5, units = "in" )
```

Model #2

```
var_imp <- varImp(rf_tune_noCCAA, scale = F)
rf2_varImp <- data.frame(
  var    = row.names(var_imp$importance),
  importance = var_imp$importance[1]
)
rf2_varImp_plot <- rf2_varImp |>
  ggplot(aes(x = Overall, xmin = 0, xmax = Overall,
             y = reorder(var, Overall))) +
  geom_linerange() + geom_point(color = "#000066") +
  geom_vline(xintercept = 0) +
  labs(y = NULL, x = "Importance")

ggsave("rf2_varImp_plot.pdf", rf2_varImp_plot, width = 9, height = 3, units = "in")
```

```
var_imp <- varImp(rf_tune, scale = F)
rf1_varImp <- data.frame(
  var    = row.names(var_imp$importance),
  importance = var_imp$importance[1]
)
rf1_varImp_plot <- rf1_varImp |>
  ggplot(aes(x = Overall, xmin = 0, xmax = Overall,
             y = reorder(var, Overall))) +
  geom_linerange() + geom_point(color = "#000066") +
  geom_vline(xintercept = 0) +
  labs(y = NULL, x = "Importance")

ggsave("rf1_varImp_plot.pdf", rf1_varImp_plot, width = 7, height = 7, units = "in")
```

```
var_imp <- varImp(rf_tune_small, scale = F)
rfsmall_varImp <- data.frame(
  var    = row.names(var_imp$importance),
  importance = var_imp$importance[1]
)
rfsmall_varImp_plot <- rfsmall_varImp |>
  ggplot(aes(x = Overall, xmin = 0, xmax = Overall,
             y = reorder(var, Overall))) +
  geom_linerange() + geom_point(color = "#000066") +
  geom_vline(xintercept = 0) +
  labs(y = NULL, x = "Importance")

ggsave("rfsmall_varImp_plot.pdf", rfsmall_varImp_plot, width = 7, height = 2, units = "in")
```

```
library(patchwork)

partial_1 <- partial(rf_tune_small, pred.var = "Eolica", plot = F) |>
```

```r
  ggplot(aes(y = yhat, x = Eolica)) + geom_line() +
  geom_point(color = "#000066") + #ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_number(scale_cut = scales::cut_long_scale()))
partial_2 <- partial(rf_tune_small, pred.var = "prec", plot = F) |>
  ggplot(aes(y = yhat, x = prec)) + geom_line() + #ylim(c(13,18)) +
  geom_point(color = "#000066")
partial_3 <- partial(rf_tune_small, pred.var = "GenTotal", plot = F) |>
  ggplot(aes(y = yhat, x = GenTotal)) +geom_line() +
  geom_point(color = "#000066") + #ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_number(scale_cut = scales::cut_long_scale()))
partial_4 <- partial(rf_tune_small, pred.var = "pop", plot = F) |>
  ggplot(aes(y = yhat, x = pop)) + geom_line() +
  geom_point(color = "#000066")+# ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_number(scale_cut = scales::cut_long_scale()))
partial_5 <- partial(rf_tune_small, pred.var = "percentage_Eólica", plot = F) |>
  ggplot(aes(y = yhat, x = percentage_Eólica)) + geom_line() +
  geom_point(color = "#000066")+ # ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_percent() )

partial_6 <- partial(rf_tune_small, pred.var = "mas", plot = F) |>
  ggplot(aes(y = yhat, x = mas)) + geom_line() +
  geom_point(color = "#000066")+ #ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_percent() )

partial_small <- (partial_1 + partial_3)/(partial_2 + partial_4) /(partial_5 + partial_6)

ggsave(partial_small, filename = "partial_small.pdf", width = 7, height = 7, units = "in" )
```

Regularized Regression

Elastic net regularization was applied to manage multicollinearity among predictors. This approach highlighted the importance of specific variables, such as population densityand total electricity generated, in predicting energy intensity.

```r
var_imp <- varImp(glmnet_tune, scale = F)
glmnet_varImp <- data.frame(
  var    = row.names(var_imp$importance),
  importance = var_imp$importance[1]
)
glmnet_varImp_plot <- glmnet_varImp |>
  ggplot(aes(x = Overall, xmin = 0, xmax = Overall,
             y = reorder(var, Overall))) +
  geom_linerange() + geom_point(color = "#000066") +
  geom_vline(xintercept = 0) +
  labs(y = NULL, x = "Importance") + theme(plot.background = element_rect(fill = "gray75"))

ggsave("glmnet_varImp_plot.pdf", glmnet_varImp_plot, width = 7, height = 5, units = "in")
```

```r
partial_1 <- partial(rf_tune_small, pred.var = "Eolica", plot = F) |>
  ggplot(aes(y = yhat, x = Eolica)) + geom_line() +
  geom_point(color = "#000066") + #ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_number(scale_cut = scales::cut_long_scale()))
partial_2 <- partial(rf_tune_small, pred.var = "prec", plot = F) |>
  ggplot(aes(y = yhat, x = prec)) + geom_line() + #ylim(c(13,18)) +
  geom_point(color = "#000066")
partial_3 <- partial(rf_tune_small, pred.var = "GenTotal", plot = F) |>
  ggplot(aes(y = yhat, x = GenTotal)) +geom_line() +
  geom_point(color = "#000066") + #ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_number(scale_cut = scales::cut_long_scale()))
partial_4 <- partial(rf_tune_small, pred.var = "pop", plot = F) |>
  ggplot(aes(y = yhat, x = pop)) + geom_line() +
  geom_point(color = "#000066")+# ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_number(scale_cut = scales::cut_long_scale()))
partial_5 <- partial(rf_tune_small, pred.var = "percentage_Eólica", plot = F) |>
  ggplot(aes(y = yhat, x = percentage_Eólica)) + geom_line() +
  geom_point(color = "#000066")+ # ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_percent() )

partial_6 <- partial(rf_tune_small, pred.var = "mas", plot = F) |>
  ggplot(aes(y = yhat, x = mas)) + geom_line() +
  geom_point(color = "#000066")+ #ylim(c(13,18))+
  scale_x_continuous(labels = scales::label_percent() )

partial_glmnet <- (partial_1 + partial_3)/(partial_2 + partial_4) /(partial_5 + partial_6)

ggsave(partial_glmnet, filename = "partial_glmnet.pdf", width = 7, height = 7, units = "in" )
```

Results

Is Month a Significant Predictor?

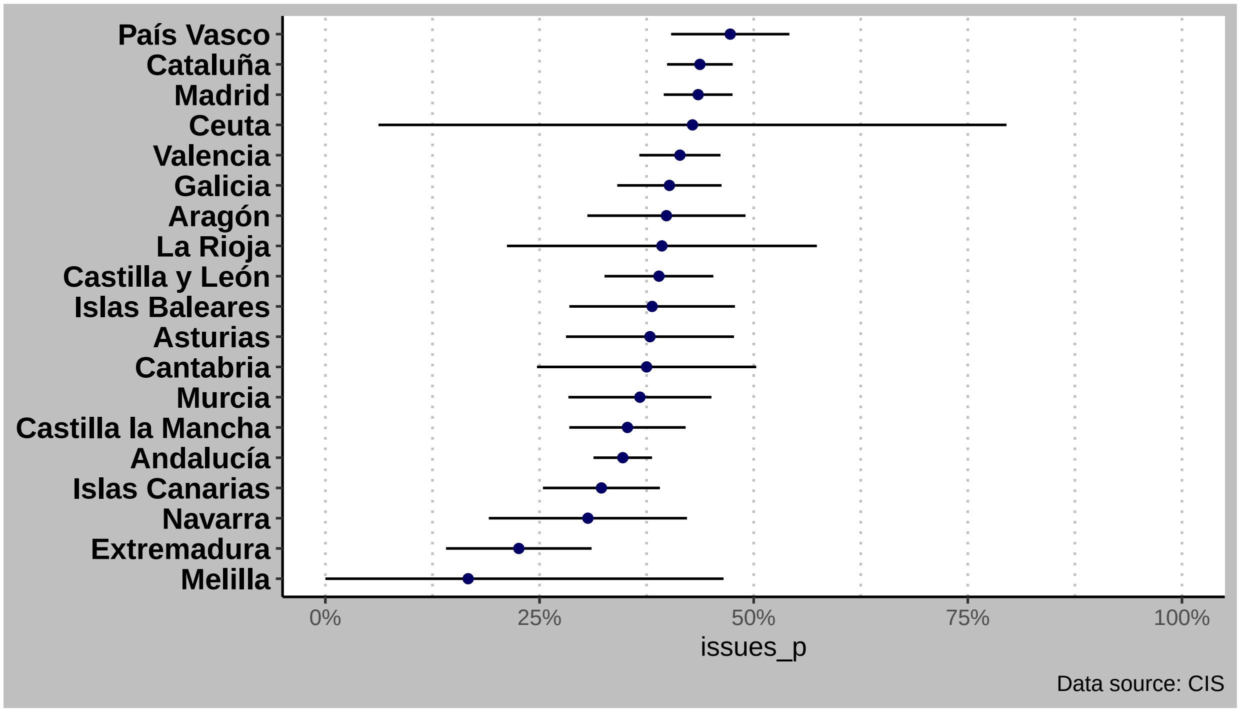
ANOVA tests were conducted to evaluate whether the month significantly predicts energy intensity. The results indicated that month alone is not a significant predictor, but its interaction with other variables, such as wind energy generation, can influence energy intensity.

CCAA as a Significant Predictor

The analysis showed that CCAA is a significant predictor of energy intensity. Mixed models incorporating CCAA provided the best fit, underscoring the importance of regional differences in energy consumption patterns.

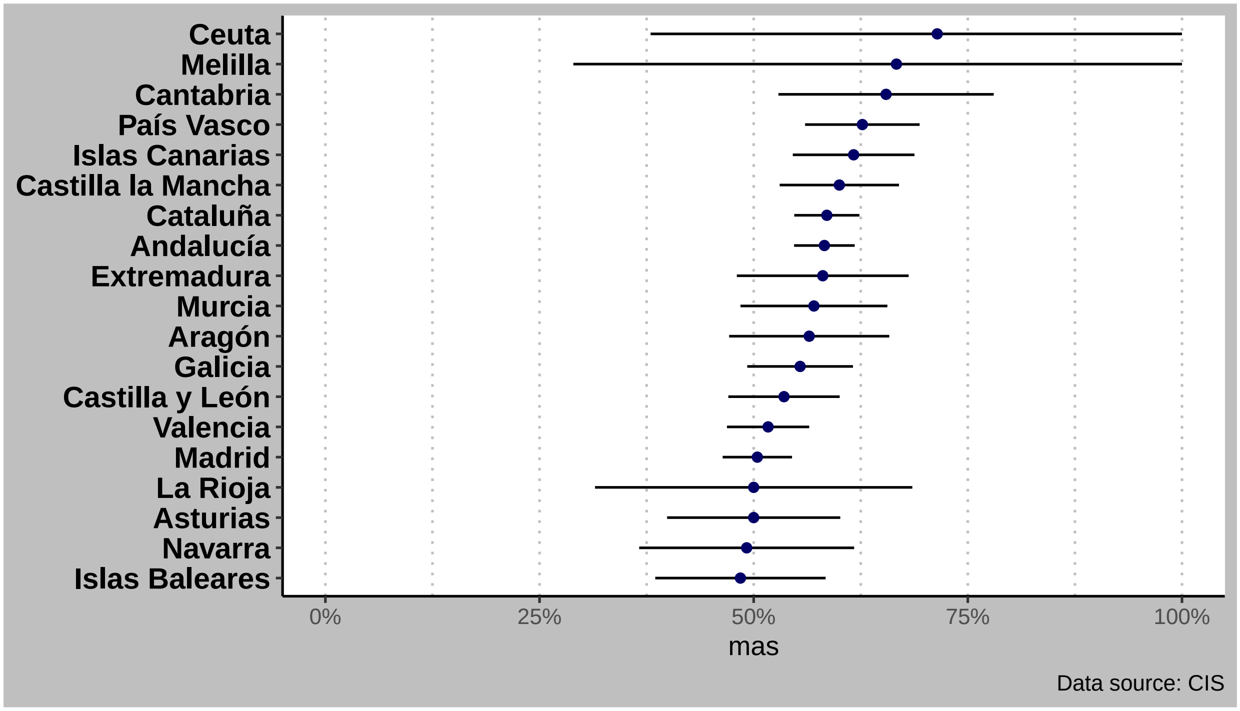
Are the survey results significant?

```r
z <-  1.96
survey |>
  select(CCAAnombre, response_p, response_sd, total.y) |>
  mutate(moe = z * sqrt(response_p * (1 - response_p) / total.y)) |> # calculate margin of error with 95
    ggplot(aes(y = reorder(CCAAnombre, response_p), x = response_p,
             xmax = moe + response_p,
             xmin = response_p - moe)) +
  geom_linerange(color = "gray0") +
  geom_point(color = "#000066") +
  labs(
    y = NULL, x = "issues_p",
    caption = "Data source: CIS"
  ) +
  scale_x_continuous(
    labels = scales::label_percent(),
    limits = c(0,1),
    oob = scales::squish
  )  +
  theme(
    axis.text.y = element_text(size = 12, face = "bold", color = "black"),
    panel.grid.major.x = element_line(color = "gray75", linetype = "dotted"),
    panel.grid.minor.x = element_line(color = "gray75", linetype = "dotted")
  )
```



```
ggsave(filename = "issues-ccaa.pdf", width = 7, height = 7, units = "in" )
```

```
z <-  1.96
survey |>
  select(CCAAnombre, mas, menos, igual, total.y) |>
  mutate(moe_mas = z * sqrt(mas * (1 - mas) / total.y)) |> # calculate margin of error with 95% conf. le
    ggplot(aes(y = reorder(CCAAnombre, mas), x = mas,
             xmax = moe_mas + mas,
             xmin = mas - moe_mas)) +
  geom_linerange(color = "gray0") +
  geom_point(color = "#000066") +
  labs(
    y = NULL,
    caption = "Data source: CIS"
  ) +
  scale_x_continuous(
    labels = scales::label_percent(),
    limits = c(0,1),
    oob = scales::squish
  )  +
  theme(
    axis.text.y = element_text(size = 12, face = "bold", color = "black"),
    panel.grid.major.x = element_line(color = "gray75", linetype = "dotted"),
    panel.grid.minor.x = element_line(color = "gray75", linetype = "dotted")
  )
```



Data source: CIS

```
ggsave(filename = "mas-ccaa.pdf", width = 7, height = 7, units = "in" )
```

Discussion

Interpretation of Results

The findings suggest that densely populated areas tend to have higher energy intensity, and increased wind energy generation is associated with increased energy intensity. This counterintuitive result may be due to inefficiencies in energy distribution or other regional factors.

Model Comparisons

Comparative analysis of different models revealed that mixed models and random forest approaches provide better predictive performance compared to simple linear models. The inclusion of regional (CCAA) and temporal (month) variables improved the models' explanatory power.

```r
comparisons_Rsquared <- stats_results |>
  mutate(Rsq  = paste0(round(100*Rsquared, 1), "%") ) |>
  filter(model != "actual_2014") |>
  ggplot(aes(xmax = Rsquared, xmin = 0,
             x = Rsquared, y = model, label = Rsq)) +
  geom_linerange(color = "#000066") +
  geom_point(color = "#000066") +
  geom_text(nudge_x = .06) +
  geom_vline(xintercept = 0) +
  labs(
    y = "Models"
  ) +
  scale_x_continuous(
    labels = scales::label_percent(),
    breaks = c(0,.2,.4, .6, .8, 1),
    limits = c(0,1)
  )

comparisons_rmse <- stats_results |>
  mutate(rmselab  = paste0(round(RMSE, 2)) ) |>
  filter(model != "actual_2014") |>
  ggplot(aes(xmax = RMSE, xmin = 0,
             x = RMSE, y = model, label = rmselab)) +
  geom_linerange(color = "#000066") +
  geom_point(color = "#000066") +
  geom_text(nudge_x = .3) +
  geom_vline(xintercept = 0) +
  labs(
    y = NULL
  )

comparisons_plot <- comparisons_Rsquared + comparisons_rmse +plot_layout(widths = c(0.6,0.4))

ggsave(comparisons_plot, filename = "comparisons-plot.pdf", width = 10, height = 4, units = "in" )
ggsave(comparisons_Rsquared, filename = "comparisons-rsq.pdf", width = 7, height = 4, units = "in" )
ggsave(comparisons_rmse, filename = "comparisons-rmse.pdf", width = 7, height = 4, units = "in" )
```

Conclusion

This study provides insights into the factors influencing energy intensity across different regions. The results emphasize the importance of considering regional and temporal variations in energy policy planning. Future research should explore the underlying causes of the observed relationships and investigate additional variables that may impact energy intensity.

References

(Include your references here, formatted according to the required citation style)

Appendices

Appendix A: Data Preprocessing Steps

Appendix B: Detailed Model Outputs

Appendix C: Additional Figures and Tables