DEGGED - Dynamic Exploration of Greenhouse Gas Emissions and its Determinants using R and Shiny

CHOONG Shi Lian Selene
Singapore Management University
slchoong.2019@mitb.smu.edu.sg

JIANG Weiling Angeline
Singapore Management University
wljiang.2019@mitb.smu.edu.sg

WONG Wei Sheng Dylan
Singapore Management University
dylan.wong.2019@mitb.smu.edu.sg

ABSTRACT

In December 2020, the European Union (EU) leaders committed to an ambitious goal of reducing greenhouse gas emissions levels by 55% by 2030 to tackle climate change. As the world's third largest emitter of greenhouse gases, it is important to determine the factors which are contributing significantly to greenhouse gas emissions. Most existing literatures focused largely on the relationship between drivers and greenhouse gas emission levels, without considering mitigation factors which also plays a role in reducing greenhouse gas emissions. Earlier research also often presented findings in static forms limiting the amount of data exploitation that can be performed. Hence, our research aims to study the relationship of both drivers and mitigation measures on greenhouse gas emission levels using Ordinary Least Squares regression and Panel Data regression. We designed and developed DEGGED, an interactive web-based dashboard, to allow policymakers and environmentalists to explore and analyse determinants of greenhouse gas emissions. DEGGED will be intuitive for non-technical users to perform fundamental data analysis and regression modeling without any coding needed from users.

1. INTRODUCTION

Global warming is expected to result in a rise of the average global temperature between 1.1 to 6.4 degree celsius over the century, if no interventions are taken to reduce the emissions of greenhouse gases [1]. Multiple drivers which include economic activities, such as electricity production, transportation and waste generation, contributes to greenhouse gases. To mitigate greenhouse gas emissions, measures like emission taxes and use of renewable energy sources were introduced. Understanding the relationship between greenhouse gas emissions and the determinants will provide greater insights on the influence of determinants on emission levels.

With large amount of environmental, social and economic statistics available on Eurostat, our aim is to provide policy makers and environmentalists with an interactive analytical tool for them to understand the influence of drivers and mitigation measures on greenhouse gas emission levels via Data Exploratory Analysis (EDA) and advanced statistical techniques namely Ordinary Least Square regression (OLS) and Panel data regression.

This paper documents our approach to design and implement a web-enabled client-based analytics tool and consists of 5 sections. Section 1 provides the general introduction of the paper. This is followed by the motivation and objectives behind our research, along with review of the existing works. Section 3 discusses on user interface design and the next section illustrates the data preparation and practical use of the interactive tool built using climate change data from Eurostat. Lastly, the paper concludes by highlighting value-add of this research and potential future work moving on from this research.

2. MOTIVATION AND OBJECTIVES

Marcotullio et al.(2013) [10] conducted exploratory data analysis on global urban greenhouse gas emissions to describe emissions by region and sector and to examine the distribution of emissions through the urban-to-rural gradient, with the descriptive findings presented in tabular forms. Kijewska & Bluszcz (2016) [9] explored including descriptive charts on the greenhouse gas emission levels over time and by countries prior to its clustering analysis. The static nature of how the data were presented in these researches restricted the amount of exploratory data analysis that users can perform.

To estimate the relationship between one or more independent variables and a dependent variable, statistical methods namely OLS regression is used to analyse at the cross-sectional level while panel data regression is employed to analyse from the longitudinal perspective. Studies from Budiono et al. (2019) [3], Grunewald & Martínez-Zarzoso (2009) [6] and Guo & Jiang (2011) [7] explored using OLS to estimate the relationship between various factors and greenhouse gas emissions. While studies from González-Sánchez & Martín-Ortega (2020) [5], Dogan & Seker (2016) [4] and Azevedo, Horta & Leal (2017) [2] investigated the determinants of greenhouse gas emissions in the European Union through panel data models.

Despite multiple past researches conducted to investigate the relationship between greenhouse gas emissions and the determinants, most research papers usually only published the final model used in their analysis and omitted the intermediate steps used to derive the final set of independent variables. Tests conducted to ensure that model assumptions were not violated were also seldom presented in the papers. In addition, tabular forms were often adopted when presenting the regression findings as well as the results of the various validation tests performed (if any).

This research was motivated by the general lack of interactive web-enabled client-based analytics tool for uncovering greenhouse gas emission patterns and influence of drivers and mitigation measures on emissions. It aims to deliver a R-Shiny app that provides interactive user-interface design to: 1) Apply fundamental data analysis to understand the factors affecting greenhouse gas emissions, including by country and time period 2) Identify main factors contributing to greenhouse gas emissions from cross-sectional view using OLS regression. Step-by-step guidance for users to detect any violation of OLS assumptions. 3) Identify main factors contributing to greenhouse gas emissions from panel data view using Panel data regression and list of validation tests available for panel analysis.

3. DESIGN FRAMEWORK

Development of the DEGGED was done on Shiny, an R package for building interactive web apps. Our design considerations when developing the application includes:

- Detailing all data preparation and modelling within R for reproducibility
- Using R packages supported on the Comprehensive R Archive Network (CRAN) for supportability
- Offering selection options and responsive visualisations for interactivity

DEGGED supports three main categories of statistical analysis, namely, (a) Exploratory data analysis, (b) OLS regression model and (b) Panel data regression model. Every interface has similar design where the left panel consists of the input variables and the output is on the right panel. The design of DEGGED follows the taxonomy of interactive dynamics from Heer & Shneiderman (2012) [8] (Data and View Specification, View Manipulation, and Process and Provenance). Detailed functionalities of each interface, including data input, parameter configuration, output views will be described in the following sub-sections. On top of these three main interfaces, DEGGED also offers users an Data Overview module to explore the data used for the application.

3.1 Exploratory Data Analysis

The EDA module consists of 4 sub-modules (Descriptive Analysis, Time Trend, Correlation matrix and Scatterplot) and aims to overcome the limitation of static visualisation in data exploration by providing users with an interactive and user-friendly display of the data, which also serve as a ground work in the variables selection for OLS and Panel data regression analysis. The plotly package in R was used for chart plotting as it has the pan and zoom features which enhances user experience.

3.1.1 Descriptive Analysis

To understand the distribution of the variables in the dataset, users can click on the Descriptive analysis module. The interface includes a slider bar for the selection of years of analysis, a drop-down list (including the option to select all or deselect all) for multiple countries selection and a drop-down list for users to select the variables of interest. (Figure. 1 - Left) Upon clicking on the Apply changes button, the summary statistics tab provides users with the measures of central tendency (Mean and Median) and measures of spread (min-max, 25th and 75th percentile) in tabular form, while the histogram tab shows the distribution of the values of the variables. By hovering over the bars in the histograms, users could read off the count of each value. (Figure. 1- Right)

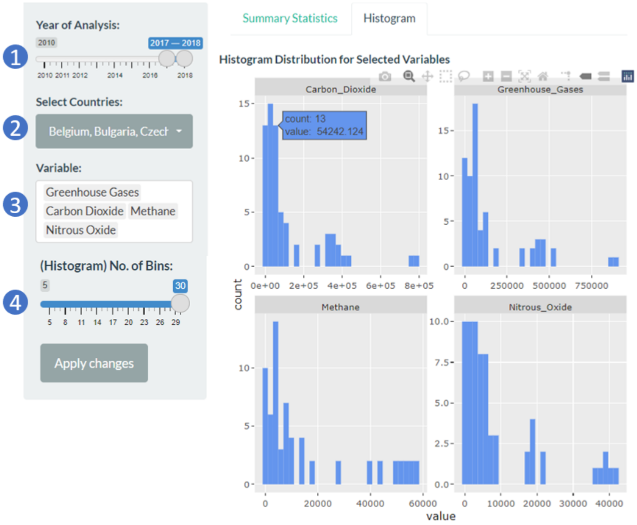


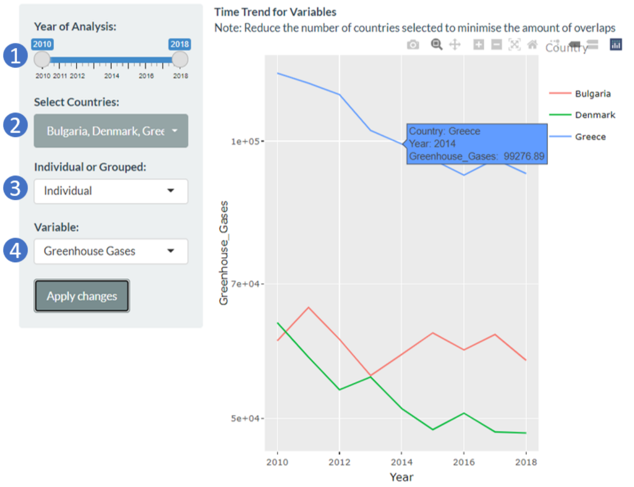
Figure. 1: User-interface for Descriptive analysis and Summary output

3.1.2 Time Trend

Users who are interested in finding patterns on how greenhouse gas emission levels or levels of drivers and mitigation factors have changed over time can click on the Time trend sub-module. Unlike the Descriptive analysis sub-module, the variable selection has been split into two drop-down lists. The Individual option allows users to compare the trend of a single variable across selected countries (Figure. 2), while the Grouped option groups the countries and displays the trend of all the variables across time (Figure. 3). Users can hover over to the data point and the actual or mean value will be displayed.

3.1.3 Correlation Matrix and Scatterplot

The Correlation matrix sub-module aims to allow users to analyse the relationship between two variables. Users can select the year of analysis from the single-select drop-down list and countries and variables selection are the same as other sub-modules where multiple selections are allowed. In addition, users have the choice on the correlation method and significance level of their preferred choice by selecting from the Type (Pearson, Spearman and Robust) and Significance level (0.01, 0.05 and 0.10) drop-down list, respectively. The correlation matrix plot or correlogram displays the pairwise correlation coefficients. The color of each variables pair tells users whether the pairs are positively (green) correlated, negatively (orange) correlated, or not (white) correlated at all. The plot also crosses out the ones that are statistically insignificant at the chosen significance level. (Figure. 4)



1 Year of Analysis: 2010 – 2018

2 Select Countries: Bulgaria, Denmark, Gree

3 Individual or Grouped: Individual

4 Variable: Greenhouse Gases

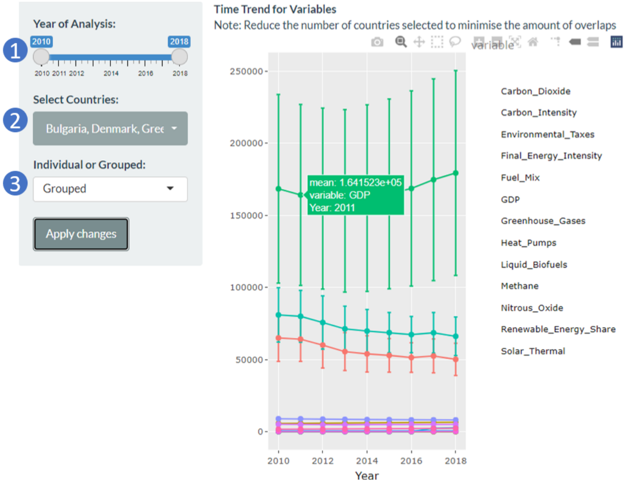
Apply changes

Time Trend for Variables

Note: Reduce the number of countries selected to minimise the amount of overlaps

Country: Greece
Year: 2014
Greenhouse_Gases: 99276.89

Figure. 2: User-interface for Time series (Individual analysis)



1 Year of Analysis: 2010 – 2018

2 Select Countries: Bulgaria, Denmark, Gree

3 Individual or Grouped: Grouped

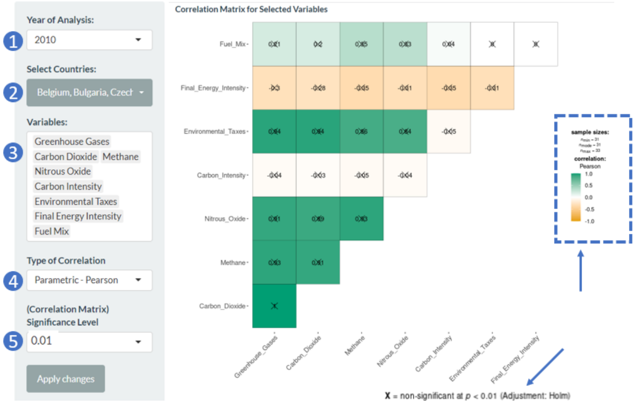
Apply changes

Time Trend for Variables

Note: Reduce the number of countries selected to minimise the amount of overlaps

mean: 1.641523e+05
variable: GDP
Year: 2011

Figure. 3: User-interface for Time series (Group analysis)



1 Year of Analysis: 2010

2 Select Countries: Belgium, Bulgaria, Czech

3 Variables: Greenhouse Gases, Carbon Dioxide, Methane, Nitrous Oxide, Carbon Intensity, Environmental Taxes, Final Energy Intensity, Fuel Mix

4 Type of Correlation: Parametric - Pearson

5 (Correlation Matrix) Significance Level: 0.01

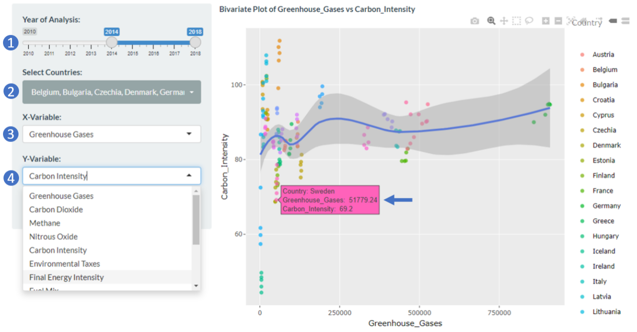
Apply changes

Correlation Matrix for Selected Variables

X = non-significant at $p < 0.01$ (Adjustment: Holm)

Figure. 4: User-interface for Correlation Matrix

For deeper analysis, users can also click on the Scatterplot sub-module as it provides users the option to plot multiple years and countries of data within a plot. In addition, the scatterplot includes a loess (locally weighted smoothing) curve to help users see the general relationship between the two selected variables. (Figure. 5)



1 Year of Analysis: 2014 – 2018

2 Select Countries: Belgium, Bulgaria, Czechia, Denmark, Germai

3 X-Variable: Greenhouse Gases

4 Y-Variable: Carbon Intensity

Bivariate Plot of Greenhouse_Gases vs Carbon_Intensity

Country: Sweden
Greenhouse_Gases: 51779.24
Carbon_Intensity: 69.2

Figure. 5: User-interface for Scatterplot

3.2 Ordinary Least Square Regression

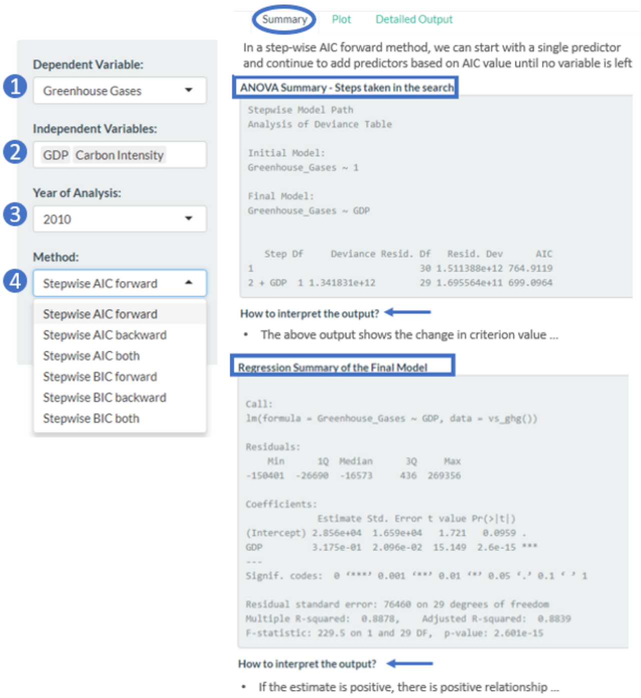
The interface of the OLS module consists of 2 sub-modules (Variable Selection and Model Selected), and provides a step-by-step guide in building OLS regression models.

3.2.1 *Variable Selection*

The first step is for users to explore and identify the best subset of independent variables to include in the model using the variable selection sub-module. Users will select the dependent and independent variables to be included in the model, and the year of analysis. Thereafter, users will click on a drop-down list consisting of different stepwise OLS methods (Stepwise AIC forward/backward/both and Stepwise BIC forward/backward/both) which are available in the MASS package in R and reactive in Shiny.(Figure. 6 - Left) Upon submitting the query, the ANOVA and regression summary will show a quick summary of the steps taken at each search and the parameter estimates of the final model. A user guide to interpreting the output was also included after each output. (Figure. 6 - Right) To have a clearer visualisation in the change of the criterion value after each search step, users can click on the Plot tab. (Figure. 7) For more detailed steps in the model ran, users can click on the Detailed output tab.

3.2.2 *Model Selected*

Based on the set of selected independent variables which the variable selection module has determined as the best fit model, users proceed to the Model selected module to obtain the parameter estimations and run the necessary OLS assumption tests. Similarly, users will select the dependent and independent variables to be included in the model, and the year of analysis. Thereafter, users can also choose from a drop-down list their preferred choice of test to run to check for any violation of OLS assumptions and check for model fit. List of tests are (i) Heteroskedasticity test (Bvensch Pagan Test; Score test; F-test), (ii) Residual diagnostics (QQ-plot; Residual vs Fitted value; Residual Histogram), (iii) Collinear diagnostics and (iv) Model fit assessment



1 Dependent Variable: Greenhouse Gases

2 Independent Variables: GDP Carbon Intensity

3 Year of Analysis: 2010

4 Method: Stepwise AIC forward

In a step-wise AIC forward method, we can start with a single predictor and continue to add predictors based on AIC value until no variable is left

ANOVA Summary - Steps taken in the search

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
Greenhouse_Gases ~ 1

Final Model:
Greenhouse_Gases ~ GDP


  Step Df     Deviance Resid. Df   Resid. Dev      AIC
1                             30 1.511388e+12 764.9119
2 + GDP  1 1.341831e+12       29 1.695564e+11 699.0964
```

How to interpret the output?

- The above output shows the change in criterion value ...

Regression Summary of the Final Model

```

Call:
lm(formula = Greenhouse_Gases ~ GDP, data = vs_ghg())

Residuals:
    Min      1Q  Median      3Q     Max
-150401  -26690  -16573     436  269356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.856e+04  1.659e+04   1.721   0.0959 .
GDP         3.175e-01  2.096e-02  15.149  2.6e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 76460 on 29 degrees of freedom
Multiple R-squared:  0.8878,    Adjusted R-squared:  0.8839
F-statistic: 229.5 on 1 and 29 DF,  p-value: 2.601e-15
```

How to interpret the output?

- If the estimate is positive, there is positive relationship ...

Figure. 6: User-interface for Variable Selection module and Summary output

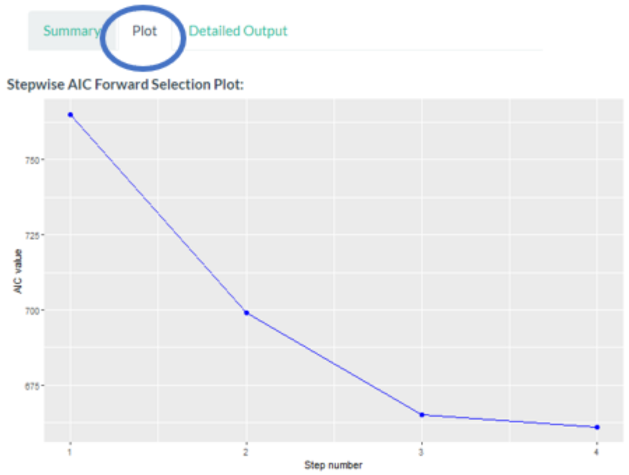
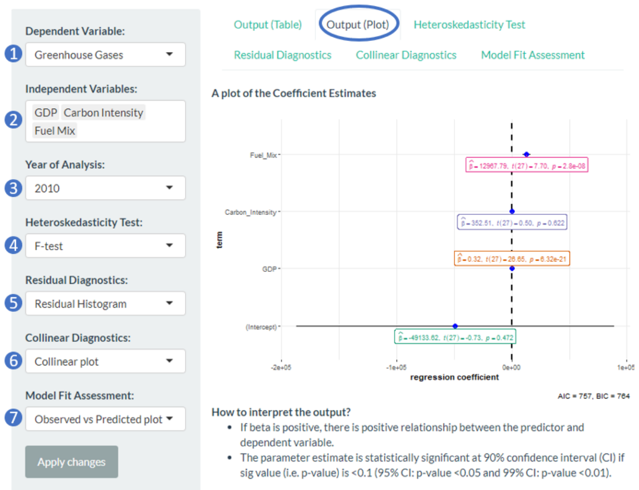


Figure. 7: Graphical summary output

(Residual Fit Spreadplot; Observed vs Predicted plot). (Figure. 8 - Left) In the output, DEGGED presents the abundant results in parameter estimation in tabular and dot plots (Figure. 8 - Right) for easy interpretation. The results from heteroskedasticity test and collinearity diagnostics are displayed in tabular form. While, results from residual diagnostics and model fit assessment will be shown using scatter plot as users will be able to easily detect any violation of the assumption by looking at the plotted points. (Figure. 9) For example, if the data is normally distributed, the points in the QQ-normal plot will lie on a straight diagonal line. The olsrr package in R was used to build the OLS assumption tests.



1 Dependent Variable: Greenhouse Gases

2 Independent Variables: GDP Carbon Intensity Fuel Mix

3 Year of Analysis: 2010

4 Heteroskedasticity Test: F-test

5 Residual Diagnostics: Residual Histogram

6 Collinear Diagnostics: Collinear plot

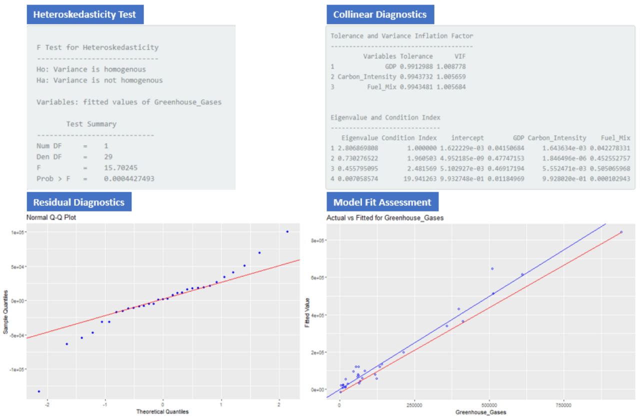
7 Model Fit Assessment: Observed vs Predicted plot

Apply changes
Output (Table)
Output (Plot)
Heteroskedasticity Test
Residual Diagnostics
Collinear Diagnostics
Model Fit Assessment
A plot of the Coefficient Estimates
Fuel_Mix
Carbon_Intensity
GDP
(Intercept)
term
$\hat{\beta} = 12967.79,\ t(27) = 7.70,\ p = 2.8e\text{-}08$
$\hat{\beta} = 352.51,\ t(27) = 0.50,\ p = 0.622$
$\hat{\beta} = 0.32,\ t(27) = 26.65,\ p = 6.32e\text{-}21$
$\hat{\beta} = -49133.62,\ t(27) = -0.73,\ p = 0.472$
-2e+05
-1e+05
0e+00
1e+05
regression coefficient
AIC = 757, BIC = 764

How to interpret the output?

- If beta is positive, there is positive relationship between the predictor and dependent variable.
- The parameter estimate is statistically significant at 90% confidence interval (CI) if sig value (i.e. p-value) is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01).

Figure. 8: User-interface for Model Selected module and Parameter estimations plot



Heteroskedasticity Test

```
F Test for Heteroskedasticity
----------------------------
Ho: Variance is homogenous
Ha: Variance is not homogenous

Variables: fitted values of Greenhouse_Gases

      Test Summary
---------------------------
Num DF     =    1
Den DF     =    29
F          =    15.70245
Prob > F   =    0.0004427493
```

Collinear Diagnostics

```
Tolerance and Variance Inflation Factor
---------------------------------------
         Variables Tolerance      VIF
1              GDP 0.9912988 1.008778
2 Carbon_Intensity 0.9943732 1.005659
3         Fuel_Mix 0.9943481 1.005684


Eigenvalue and Condition Index
------------------------------
   Eigenvalue Condition Index    intercept        GDP Carbon_Intensity    Fuel_Mix
1 2.806869808        1.000000 1.622229e-03 0.04150684     1.643634e-03 0.042278331
2 0.730276522        1.960503 4.952185e-09 0.47747153     1.846496e-06 0.452552757
3 0.455795095        2.481569 5.102927e-03 0.46917194     5.552471e-03 0.505065968
4 0.007058574       19.941263 9.932748e-01 0.01184969     9.928020e-01 0.000102943
```

Residual Diagnostics

Model Fit Assessment

Figure. 9: OLS Assumption tests

3.3 Panel Data Regression

Panel data regression extends beyond OLS to consider the longitudinal aspect of the data, where users can examine the variations along both the individual and time dimensions. The complexity involved in constructing panel models and ensuring its robustness will be addressed by DEGGED which simplifies the process into 2 sub-modules (Model Building and Validation Tests), and provides a step-by-step guide in building and evaluating panel regression models. As OLS is explored as a separate module in the earlier section, the panel regression module only covers fixed effects and random effects models.

The fixed effects model considers individual differences over time, and assumes correlation between the individuals' error term and predictor variables, with no correlation between time-invariant and individuals' other characteristics. The random effects model considers unique errors to be uncorrelated with the regressors, and assumes that the variations across individuals to be random and uncorrelated with the predictor variables. The plm package in R was used to build the panel data models.

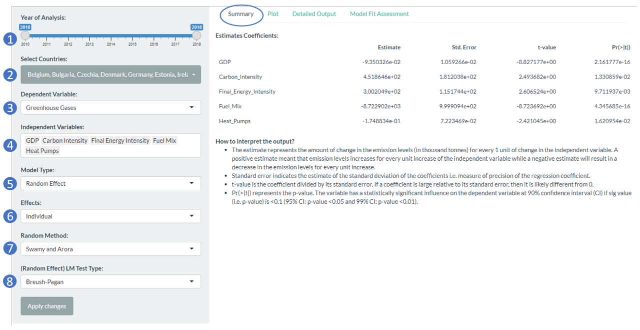
3.3.1 Model Building

Apart from the dynamic time and country dimensions, users can select the independent and dependent variables of interest to be included in the model. Users can subsequently decide the type of panel data models (Fixed effect model and Random effect model) as well as the type of effect (Individual, Time and Two-ways). For random effect model, the application allows for four different estimators of the parameter, namely Swamy and Arora, Wallace and Hussain, Amemiya and Nerlove. Users can also select the type of Lagrange Multipler test (Breush-Pagan, Honda and King & Wu) to perform for the subsequent Model Fit Assessment. (Figure. 10 - Left) Running the selected model will produce a coefficient estimate table (Figure. 10 - Right) along with a dot-and-whisker plot of the regression coefficient estimates (Figure. 11) for clearer visualisation. For more statistical details, users can also refer to the Detailed output tab.

Upon running the selected panel regression model, users will also be given the Model Fit Assessment to assess which of panel model (pooled OLS, fixed effect or random effect model) is most suitable for the selected variables and data used. To assess if fixed or random effect model should be used, DEGGED conducts the Hausman Test with the null hypothesis being the preferred model to be random effects. To decide if fixed effect model should be used over pooled OLS, Chow Test of Poolability will be performed with the null hypothesis being the pooled effect model. Lagrange Multiplier Test, with null hypothesis being no panel effect, is used to determine if pooled OLS or random effect model should be used. Relevant tests will be shown based on the Model Type selected. Outputs from each of the model selection tests are presented in tabular forms due to the limitations of clutteredness when approaching the assessments using visual representations. (Figure. 12)

3.3.2 *Tests for Assumptions*

To ensure the robustness of the model, additional diagnostic tests were performed to ensure that the dataset complies with the assumptions of panel regression model. Users are offered the option to choose from a drop-down list their preferred choice of test to check for any violation of panel regression assumptions (Figure. 13 - Left). The list of diagnostic tests included in DEGGED are (i) Test of Normality (Residual QQ Plot; Residual vs Fitted Value Plot; Residual Histogram), (ii) Test of Serial Correlation (Unobserved Effect Test; Locally Robust Tests; Breusch-Godfrey/Wooldridge Test) and (iii) Test for Heteroskedasticity (Breusch-Pagan Test; Residual vs Fitted Value Plot). DEGGED presents the results from the various tests in either tabular form or charts depending on outputs (Figure. 13 - Right). Intepretation reference is included in the application to help guide users interpret the results.



Year of Analysis: 2010 – 2018

Select Countries: Belgium, Bulgaria, Czechia, Denmark, Germany, Estonia, Irel

Dependent Variable: Greenhouse Gases

Independent Variables: GDP, Carbon Intensity, Final Energy Intensity, Fuel Mix, Heat Pumps

Model Type: Random Effect

Effects: Individual

Random Method: Swamy and Arora

(Random Effect) LM Test Type: Breush-Pagan

Estimates Coefficients:

	Estimate	Std. Error	t-value	Pr(>|t|)
GDP	-9.350326e-02	1.059266e-02	-8.827177e+00	2.161777e-16
Carbon_Intensity	4.518646e+02	1.812038e+02	2.493682e+00	1.330859e-02
Final_Energy_Intensity	3.002049e+02	1.151744e+02	2.606524e+00	9.711937e-03
Fuel_Mix	-8.722902e+03	9.999094e+02	-8.723692e+00	4.345685e-16
Heat_Pumps	-1.748834e-01	7.223469e-02	-2.421045e+00	1.620954e-02

How to interpret the output?

- The estimate represents the amount of change in the emission levels (in thousand tonnes) for every 1 unit of change in the independent variable. A positive estimate meant that emission levels increases for every unit increase of the independent variable while a negative estimate will result in a decrease in the emission levels for every unit increase.
- Standard error indicates the estimate of the standard deviation of the coefficients i.e. measure of precision of the regression coefficient.
- t-value is the coefficient divided by its standard error. If a coefficient is large relative to its standard error, then it is likely different from 0.
- Pr(>|t|) represents the p-value. The variable has a statistically significant influence on the dependent variable at 90% confidence interval (CI) if sig value (i.e. p-value) is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01).

Figure. 10: User-interface for Model Building module and Summary output

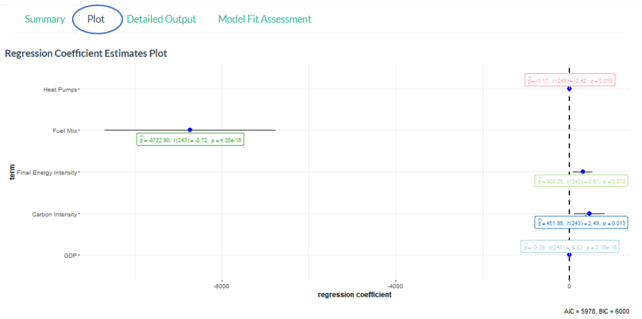
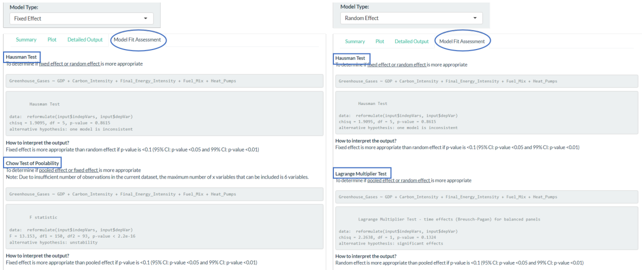


Figure. 11: Graphical output of coefficient estimates



Model Type: Fixed Effect

Hausman Test

To determine if fixed effect or random effect is more appropriate

```
Greenhouse_Gases ~ GDP + Carbon_Intensity + Final_Energy_Intensity + Fuel_Mix + Heat_Pumps
```

```
        Hausman Test

data:  reformulate(input$indepVars, input$depVar)
chisq = 1.9095, df = 5, p-value = 0.8615
alternative hypothesis: one model is inconsistent
```

How to interpret the output?
Fixed effect is more appropriate than random effect if p-value is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01)

Chow Test of Poolability

To determine if pooled effect or fixed effect is more appropriate
Note: Due to insufficient number of observations in the current dataset, the maximum number of x variables that can be included is 6 variables.

```
Greenhouse_Gases ~ GDP + Carbon_Intensity + Final_Energy_Intensity + Fuel_Mix + Heat_Pumps
```

```
        F statistic

data:  reformulate(input$indepVars, input$depVar)
F = 13.153, df1 = 150, df2 = 93, p-value < 2.2e-16
alternative hypothesis: unstability
```

How to interpret the output?
Fixed effect is more appropriate than pooled effect if p-value is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01)

Model Type: Random Effect

Hausman Test

To determine if fixed effect or random effect is more appropriate

```
Greenhouse_Gases ~ GDP + Carbon_Intensity + Final_Energy_Intensity + Fuel_Mix + Heat_Pumps
```

```
        Hausman Test

data:  reformulate(input$indepVars, input$depVar)
chisq = 1.9095, df = 5, p-value = 0.8615
alternative hypothesis: one model is inconsistent
```

How to interpret the output?
Fixed effect is more appropriate than random effect if p-value is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01)

Lagrange Multiplier Test

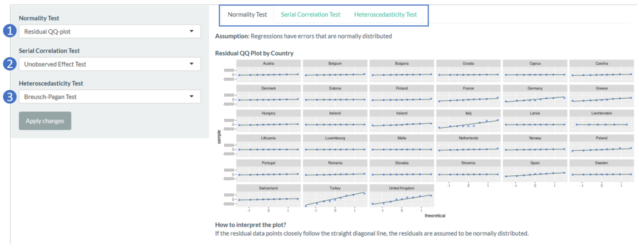
To determine if pooled effect or random effect is more appropriate

```
Greenhouse_Gases ~ GDP + Carbon_Intensity + Final_Energy_Intensity + Fuel_Mix + Heat_Pumps
```

```
        Lagrange Multiplier Test - time effects (Breusch-Pagan) for balanced panels

data:  reformulate(input$indepVars, input$depVar)
chisq = 2.2638, df = 1, p-value = 0.1324
alternative hypothesis: significant effects
```

How to interpret the output?
Random effect is more appropriate than pooled effect if p-value is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01)

Figure. 12: Model Fit Assessment for constructed panel model



Normality Test: Residual QQ-plot

Serial Correlation Test: Unobserved Effect Test

Heteroscedasticity Test: Breusch-Pagan Test

Assumption: Regressions have errors that are normally distributed

Residual QQ Plot by Country

How to interpret the plot?
If the residual data points closely follow the straight diagonal line, the residuals are assumed to be normally distributed.

Figure. 13: User interface for Tests of Assumption module

4. USE CASE: GHG EMISSIONS IN EU

In December 2020, the European Union (EU) leaders committed to an ambitious goal of reducing greenhouse gases by 55% by 2030 to tackle climate change. As the world's third biggest emitter of greenhouse gases, it is important to determine the factors which are contributing significantly to greenhouse gas emissions.

A study conducted by González-Sánchez & Martín-Ortega (2020) [5] on the determinants of greenhouse gas emissions growth in Europe considered majority of the driver factors available on Eurostat database and found that GDP and final energy intensity are the main drivers for the reduction of greenhouse gas emissions in Europe. However, mitigation factors which also plays a part in reducing greenhouse gas emissions were not considered in their study. Hence, our research further explored the Eurostat database and extended the list with mitigation factors which could potentially play an important role in the reduction of greenhouse gas emissions in Europe.

The data used for our research contains the overall greenhouse gas emissions, excluding Land Use and Land Use Change and Forest (LULUCF) and the emission levels of common greenhouse gases (i.e. carbon dioxide, methane and nitrous oxide) from 33 European countries between the period 2010 to 2018. Data from four drivers (Gross Domestic Product, Final energy intensity, Fuel mix and Carbon intensity) and five mitigation factors (Renewable energy, Environmental taxes by economic activity, Liquid biofuels production capacities, Solar thermal collectors' surface and Heat pumps - technical characteristics by technologies) were also included.

As the downloaded datasets from Eurostat database had irrelevant rows and columns, data preparation steps were performed to tidy the data prior to combining into one dataset for our reserach use. Additional computation were also performed to obtain the final energy intensity as data for the indicator was not readily available. While Eurostat has distinguished the production capacities by different liquid biofuel types and the installed thermal capacity by different heat pump technologies, our research will only be analyzing these factors at the collective overall level. Hence, aggregation was performed to get the total value of the respective factors for each country at each year. The combined dataset was finally transposed to obtain a country-year record level data for analyzing.

Our assessment showed that DEGGED has several advantages over traditional exploratory and regression analysis. Firstly, upon detecting any unusual patterns from the overall view at the data exploratory step, users can pan and zoom to focus on specific data points with more information displayed by hovering over to the data point, without having the need to filter through the raw dataset. (Figure. 14) The pan and zoom features are from plotly package in R.

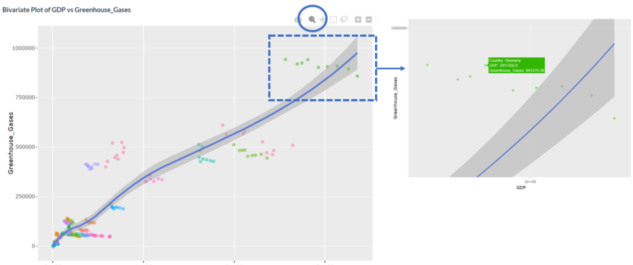
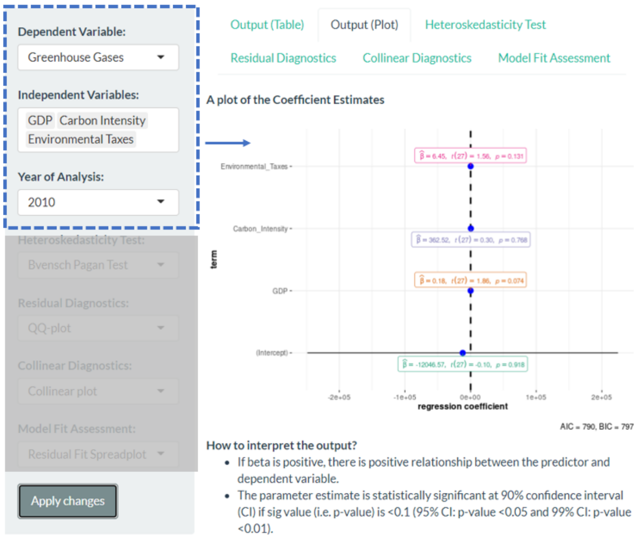


Figure. 14: Data exploration using pan and zoom features

Secondly, the interactive features available in the OLS (Figure. 15) and Panel Regression (Figure. 16) modules enable users to choose their list of independent and dependent variables and run their preferred model without any coding or having to code multiple code chunks to cater for different specifications. The regression output, both tabular and graphical format, are almost updated instantaneously. DEGGED also has a pre-loaded list of necessary assumption tests to detect any violation of the model assumptions each time a model is ran. (Figure. 17) (Figure. 18) To give users more options, users can also select their preferred test from the drop-down list of each assumption test. For example, users can choose between Breusch Pagan test, Score-test or F-test to detect heteroskedasticity problem. As a value-add,



Dependent Variable: Greenhouse Gases

Independent Variables: GDP Carbon Intensity Environmental Taxes

Year of Analysis: 2010

Heteroskedasticity Test: Bvensch Pagan Test

Residual Diagnostics: QQ-plot

Collinear Diagnostics: Collinear plot

Model Fit Assessment: Residual Fit Spreadplot

A plot of the Coefficient Estimates

$\hat{\beta} = 6.45,\ t(27) = 1.56,\ p = 0.131$

$\hat{\beta} = 362.52,\ t(27) = 0.30,\ p = 0.768$

$\hat{\beta} = 0.18,\ t(27) = 1.86,\ p = 0.074$

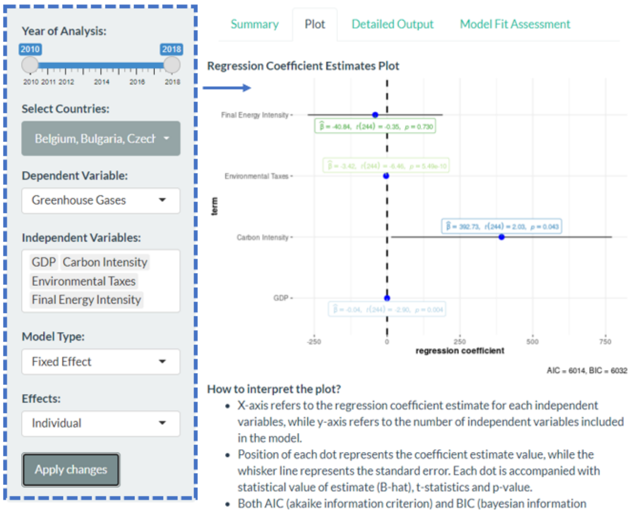
$\hat{\beta} = -12046.57,\ t(27) = -0.10,\ p = 0.918$

AIC = 790, BIC = 797

How to interpret the output?

- If beta is positive, there is positive relationship between the predictor and dependent variable.
- The parameter estimate is statistically significant at 90% confidence interval (CI) if sig value (i.e. p-value) is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01).

Figure. 15: Interactive parameter estimations for OLS



Year of Analysis: 2010 2018

Select Countries: Belgium, Bulgaria, Czech

Dependent Variable: Greenhouse Gases

Independent Variables: GDP Carbon Intensity Environmental Taxes Final Energy Intensity

Model Type: Fixed Effect

Effects: Individual

Regression Coefficient Estimates Plot

$\hat{\beta} = -40.84,\ t(244) = -0.35,\ p = 0.730$

$\hat{\beta} = -3.42,\ t(244) = -6.46,\ p = 5.49e\text{-}10$

$\hat{\beta} = 392.73,\ t(244) = 2.03,\ p = 0.043$

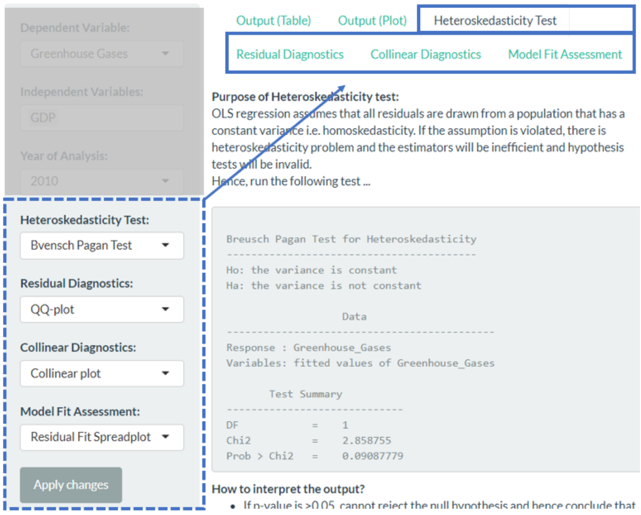
$\hat{\beta} = -0.04,\ t(244) = -2.90,\ p = 0.004$

AIC = 6014, BIC = 6032

How to interpret the plot?

- X-axis refers to the regression coefficient estimate for each independent variables, while y-axis refers to the number of independent variables included in the model.
- Position of each dot represents the coefficient estimate value, while the whisker line represents the standard error. Each dot is accompanied with statistical value of estimate (B-hat), t-statistics and p-value.
- Both AIC (akaike information criterion) and BIC (bayesian information

Figure. 16: Interactive parameter estimations for Panel



Dependent Variable: Greenhouse Gases

Independent Variables: GDP

Year of Analysis: 2010

Heteroskedasticity Test: Bvensch Pagan Test

Residual Diagnostics: QQ-plot

Collinear Diagnostics: Collinear plot

Model Fit Assessment: Residual Fit Spreadplot

Purpose of Heteroskedasticity test:
OLS regression assumes that all residuals are drawn from a population that has a constant variance i.e. homoskedasticity. If the assumption is violated, there is heteroskedasticity problem and the estimators will be inefficient and hypothesis tests will be invalid.
Hence, run the following test ...

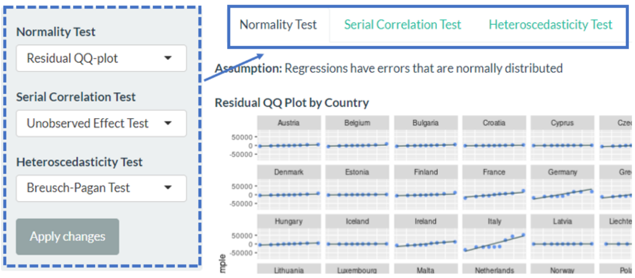
```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

                  Data
-------------------------------------------
Response : Greenhouse_Gases
Variables: fitted values of Greenhouse_Gases

       Test Summary
-----------------------------
DF            =    1
Chi2          =    2.858755
Prob > Chi2   =    0.09087779
```

How to interpret the output?

- If p-value is >0.05 cannot reject the null hypothesis and hence conclude that

Figure. 17: Interactive assumption tests for OLS

DEGGED also includes interactive notes at the bottom of each output to guide and explain how each output should be interpreted. (Figure. 19)

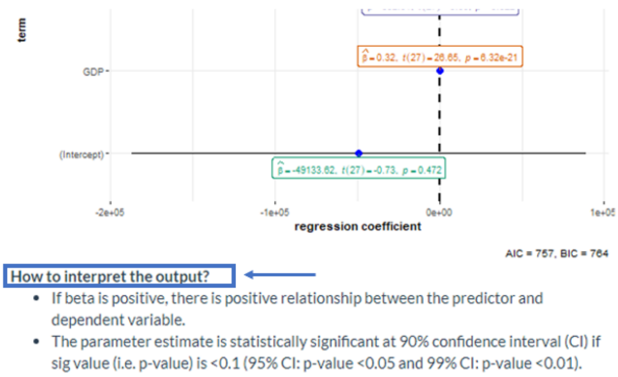

Normality Test: Residual QQ-plot

Serial Correlation Test: Unobserved Effect Test

Heteroscedasticity Test: Breusch-Pagan Test

Assumption: Regressions have errors that are normally distributed

Figure. 18: Interactive asumption tests for Panel


term
GDP
(Intercept)
$\hat{\beta} = 0.32,\ t(27) = 26.65,\ p = 6.32\text{e-}21$
$\hat{\beta} = -49133.62,\ t(27) = -0.73,\ p = 0.472$
-2e+05
-1e+05
0e+00
1e+05
regression coefficient
AIC = 757, BIC = 764

How to interpret the output?

- If beta is positive, there is positive relationship between the predictor and dependent variable.
- The parameter estimate is statistically significant at 90% confidence interval (CI) if sig value (i.e. p-value) is <0.1 (95% CI: p-value <0.05 and 99% CI: p-value <0.01).

Figure. 19: Interactive interpretation guide

5. CONCLUSION & FUTURE WORK

The demonstration of DEGGED clearly shows the benefits that policy makers and environmentalists, with or without technical background, can reap as they will be able to perform on the fly data exploratory analysis and advanced statistical techniques (OLS and Panel data regression) to gain quick and valuable insights for policy making. Nonetheless, instead of using MASS package in R which outputs only AIC criterion value at the stepwise modeling step, future work could be explored to modify the olsrr package in R for it to incorporate reactive dataset and hence able to output a panel of fit criterion (R-square, Adjusted R-square, Mallow's Cp, AIC, SBIC and SBC). In addition, the scope of DEGGED could be extended to cover greenhouse gas emissions in other countries such as U.S. as the U.S. aims to cut greenhouse gas emissions in half by 2030 as part of its new commitment to the Paris climate agreement.

6. ACKNOWLEDGEMENT

The authors wish to thank Professor Kam Tin Seong of School of Computing and Information System at Singapore Management University for his support and guidance throughout this project.

References

[1] Amanatidis, G. 2020. Combating climate change. European Parliament.

[2] Azevedo, I. et al. 2017. Analysis of the relationship between local climate change mitigation actions and greenhouse gas emissions – Empirical insights. Energy policy. 111, (Dec. 2017), 204–213.

[3] Budiono, R. et al. 2019. Modeling and analysis of CO2 emissions in million tons of sectoral greenhouse gases in Indonesia. IOP Conference Series. Materials Science and Engineering. 621, 1 (Oct. 2019).

[4] Dogan, E. et al. 2016. An investigation on the determinants of carbon emissions for OECD countries - empirical evidence from panel models robust to heterogeneity and cross-sectional dependence. Environmental science and pollution research international. 23, 14 (Jul. 2016), 14646–14655.

[5] González-Sánchez, M. and Martín-Ortega, J.L. 2020. Greenhouse Gas Emissions Growth in Europe - A Comparative Analysis of Determinants. Sustainability (Basel, Switzerland). 12, 3 (Jan. 2020), 1012.

[6] Grunewald, N. and Martínez-Zarzoso, I. 2010. Driving Factors of Carbon Dioxide Emissions and the Impact from Kyoto Protocol.

[7] Guo, H. and Jiang, Y. 2011. The Relationship between CO2 Emissions, Economic Scale, Technology, Income and Population in China. Procedia Environmental Sciences. 11 (2011), 1183–1188.

[8] Heer, J. and Shneiderman, B. 2012. Interactive Dynamics for Visual Analysis. ACM Queue. 55, 4 (Feb. 2012), 45–54.

[9] Kijewska, A. and Bluszcz, A. 2016. Analysis of greenhouse gas emissions in the European Union member states with the use of an agglomeration algorithm. Journal of Sustainable Mining. 15, 4 (Nov. 2016), 133–142.

[10] Marcotullio, P.J. et al. 2013. The geography of global urban greenhouse gas emissions - an exploratory analysis. Climatic change. 121, 4 (Dec. 2013), 621–634.