Early estimates of energy balance statistics using machine learning

Anneleen Goyens
Bart Buelens

15 June 2020

VITO NV
Boeretang 200 - 2400 MOL - BELGIE
Tel. + 32 14 33 55 11 - Fax + 32 14 33 55 99
vito@vito.be - www.vito.be

BTW BE-0244.195.916 RPR (Turnhout)
Bank 375-1117354-90 ING
BE34 3751 1173 5490 - BBRUBEBB

SUMMARY

We report on a study conducted at VITO in the context of the UNECE Machine Learning project, investigating the use of machine learning methods for official statistics. Our study is based on the Energy Balance, a collection of statistics on production and consumption of energy. These are annual statistics and become available each year in September or October, reporting on the preceding year. In this study we research possibilities to publish earlier, by imputing missing data cells with predictions. Predictions are made for data that are not available yet, using other data sources as predictors, such as economic and weather variables in our case.

We composed a data set containing quarterly data for a selection of target variables and predictors, for the years 2000 through 2019. We split the data in training and test sets and apply prediction methods including linear regression, ridge and lasso regression, random forests and neural networks. We compare predictions by these methods to predictions by a baseline method that simply uses the observed value of the same quarter in the previous year as the prediction for the current year. The results are mixed, in that the baseline method outperforms the machine learning methods for three of the eight target variables under consideration. For the other five, machine learning methods do improve on the baseline method. No single method stands out as working best. Neural networks and random forests perform well in several cases, but also the linear model works fine for one variable. In several cases, an ensemble estimator consisting of averages of the other methods beats each of the individual methods, and hence looks promising.

This work is preliminary and conducted with limited resources. We could consider a wider range of methods, and tune the hyper-parameters of some of the methods better. An important improvement would be adding estimates of uncertainty, for example by using bootstrap procedures. Not only would estimates of uncertainty be relevant in their own right, they would also improve the ensemble estimator which could include the uncertainties in its averaging procedure.

Our study shows that there are no obvious quick wins to be made, and that the uptake of machine learning methods in standard procedures requires substantial and continued effort and commitment. Our results should be seen as a first exploration of possibilities in the context of the Energy Balance. This work can be used as a basis or reference for future research on this topic at VITO or elsewhere. We are making our data and code publicly available to stimulate continued research and to enable reproducibility of our work.

Data used in this research: https://doi.org/10.5281/zenodo.3596694
Code implementing the analyses: https://github.com/VITObelgium/energy-balance-ml

TABLE OF CONTENTS

CHAPTER 1 BACKGROUND

1.1. BACKGROUND AND WHY AND HOW THIS STUDY WAS INITIATED

Countries all over the world are struggling to meet climate change targets, and Belgium is no exception. Belgium has three regions (Flanders, Wallonia, Brussels Region) and governments both at the regional and federal level, each with their own plans and objectives. This fragmentation is often criticized in the media since it leads to difficult decision making. There is a need for cooperation based on adaptivity and agility to attain the sustainable development goals. Energy statistics can help understand the situation better and make better decisions for the future.

Every year, an energy balance report is produced for the Flemish region by VITO, an independent Flemish research and technology organization in the area of cleantech and sustainable development. The mission of VITO is to accelerate the transition to a sustainable world. The annual energy balance report is organized by economic sector and by energy source, and provides a general overview of energy use and the production of fossil fuels and renewable energy sources in Flanders. It charts the evolution of the Flemish energy flows since 1990. The report is submitted to the Federal Government in the form of a response to an annual questionnaire drawn up by the IEA (International Energy Agency) and Eurostat, as are the reports of the other two regions (Wallonia and Brussels Region). The Federal Government combines the data from the three reports, making a number of additional analyses of its own, before submitting it to Eurostat in order to meet European reporting requirements. The report is thus a key instrument for stakeholders such as the government, the European commission and other political players.

The energy balance report is important to monitor the share of renewable energy in the energy mix and to gain insight into Belgium's energy dependence and security of supply. The statistics in the energy balance report can be used to evaluate policy actions in energy and environment and to quantify the impacts of political measures. They also serve as a basis for other calculations, models and studies both inside and outside VITO.

1.2. ENERGY BALANCE: OUTDATED PROCESS AND MISSING DATA

The yearly process of creating an energy balance report at VITO is very time consuming and outdated, with old-fashioned working methods, such as large and complex Excel sheets. The use of these Excels often leads to data errors with snowball effects, since a change in one Excel cell can affect many other parts of the process. A second issue is that it is difficult to predict when the data sources will become available during the first half of the year and what their level of completeness will be. This makes it difficult to plan and almost always leads to delays in producing the report. For many years, there has been no detailed investigation into new techniques to improve and speed up the process of reporting.

An increasing number of organizations are producing statistics in a more timely and accessible manner. Interest in the use of Machine Learning (ML) for official statistics is growing at a fast pace (Hassani et al, 2014; Daas et al, 2015). Opportunities that this trend could bring are discussed in the

UNECE position paper (UNECE Machine Learning Team, 2018). VITO does not wish to lag behind and is implementing measures and taking steps to stay ahead of the competition. It seems indispensable to look into less work intensive methods offered by modern ML techniques. Although ML seems promising, few guidelines and good practices have been developed to date and some issues still have to be solved. For example, one challenge is to use consistent definitions and units to safeguard the quality, consistency and compatibility of energy data in ML applications. A second is to ensure a match in the approaches to forecasting energy supply and demand. Thirdly, since ML appears to be a complex technique, it is important to guarantee transparency and make the technique defensible to energy balance stakeholders. These are all areas in which VITO – and other users of ML, of course – could benefit from greater insight.

One of the main advantages of using Machine Learning would be the shorter production process for the final report. Many stakeholders are showing interest in early estimates and are accelerating deadlines to deliver the energy statistics. Demand for early estimates of the energy balance, for example earlier than six months after the end of the reference period, is growing. For VITO, delivering early estimates is the goal for the research reported here. Early and more accurate estimates could provide added value for VITO and for the clients. However, an investment will have to be made to investigate new techniques and the return on this investment will have to be justified carefully.

1.3. THE ML STUDY: PREDICTIONS WITH NON-ENERGY RELATED INDICATORS

When we were asked to participate in the UNECE ML project, we seized the opportunity. For this project we started a collaboration between TEEM (the department within VITO responsible for delivering the energy balance report) and the data science hub of VITO. We started this study in order to demonstrate what can be achieved with ML and how we can implement it in a responsible manner so we can make it defensible to our clients. This research was an opportunity to start using ML to improve prediction tasks and to share knowledge internally and externally.

This ML study is an imputation project with time dependencies. We aim to fill in energy consumption per sector/carrier for the energy balance report of the reference year with forecasts based on over 50 economic indicators and two weather indicators. At present, the data for different sectors and energy carriers is received at different points of the year and in different degrees of completeness. In order to deliver early estimates, the missing data must be imputed with forecasts. The economic indicators and weather indicators, in contrast, are available at set times earlier in the year. By using these non-energy related indicators, we can overcome the issue of delayed data and take steps towards producing earlier estimates.

1.4. DATA AND SCOPE OF THE STUDY

During the study, we opted to begin with a yearly dataset for Flanders from 1995 until 2012. We moved to a quarterly dataset of Belgium to have more data points (#77): quarterly data from 2000 Q1 until 2019 Q1. The variables under consideration are listed in Table 1. Three types of variables are distinguished:

- Target variables: Monthly Electricity Supplied – IEA (https://www.iea.org/reports/monthly-oecd-electricity-statistics)
- Economic indicators: GDP, GVA,… - (www.statistiekvlaanderen, stat.nbb.be)
- Weather indicators: Degree days and sun spots - (www.gas.be/nl/graaddagen, http://www.sidc.be/silso/datafiles

The economic and weather variables together are used as predictors or auxiliary variables. The scope of the energy variable we want to predict is the electricity supplied in Belgium. The electricity supplied is only a part of the total energy supplied in Belgium.

Table 1. Description of variables, corresponding to the data set publicly available, see section 5.1. For the target variables, the names used in the code and in the results section are added in italics in brackets.

Target variables	**Description**
+ Combustible Fuels GWh (*EnrgCombustibleFuels*)	Production from fossil fuels (primary coal, coal products, peat and peat products, oil shale and oil sands, crude oil, NGL, oil products, natural gas) and combustible renewables and wastes (solid biofuels, biogases, liquid biofuels, industrial and municipal waste).
+ Nuclear GWh (*EnrgNuclearNuclear*)	Electricity produced using heat generated from nuclear fission.
+ Hydro GWh (*EnrgHydroHydro*)	Net generation from hydro facilities including pumped storage production.
+ Geothermal/Other GWh (*EnrgGeothermalOther*)	Generation from geothermal, solar photovoltaic, solar thermal, wind, tide, wave, ocean and other non-combustible sources.
=Indigenous Production GWh (*EnrgIndigenousProd*)	The sum of electricity production by energy source.
+ Imports GWh (*EnrgImportsImports*)	Amounts of electricity that have crossed political boundaries of the country,
- Exports GWh (*EnrgExportsExports*)	Amounts of electricity that have crossed political boundaries of the country,
= Electricity Supplied GWh (*EnrgElectricitySupplied*)	Indigenous production + Imports - Exports. It includes transmission and distribution losses.
Economic variables	
Gross domestic income	The Gross Domestic Income (GDI) is the total income received by all sectors of an economy within a state.
Balance of primary incomes receivable from/payable to the rest of the world	
Gross national income	
Consumption of fixed capital	
Net national income	
Balance of current transfers receivable from/payable to the rest of the world	
Net national disposable income	
National final consumption	
Net national saving	
Gross fixed capital formation	
Changes in stocks	

Balance of capital transfers receivable from/payable to the rest of the world	
Net lending (+) or borrowing (-)	
Final consumption expenditure of households in Belgium (inland) - estimates in prices	
Durable goods - estimates in prices	
Others - estimates in prices	
Final consumption expenditure of households in Belgium (inland) - estimates in volume	
Durable goods - estimates in volume	
Others - estimates in volume	
Imports Total	
Exports Total	
GVA Total economy	The gross value added (**GVA**) is the measure of the value of goods and services produced in an area, industry or sector of an economy
GVA Agriculture and forestry and fishing	
GVA Industry except construction	
GVA Construction	
GVA Services Wholesale and retail trade/ transport/ accomodation and food service activities	
GVA Services Information and communication	
GVA Services Financial and insurance activities	
GVA Services Real estate activities	
GVA Services Professional and scientific and technical activities plus administrative and support service activities	
GVA Services Other community and social and personal service activities	
GVA Services Human health and social work activities	
GVA Services Arts and entertainment and recreation/ other service activities/ activities of household and extra-territorial organizations and bodies	
Total economy taxes less subsidies on products	
Gross domestic product	
GDP Private Final consumption	Gross domestic product (**GDP**) is a monetary measure of the market value of all the final goods and services produced in a specific time period.
GDP Final consumption expenditure of general government	
GDP Gross fixed capital formation	
GDP Gross fixed capital formation by enterprises, self-employed workers and non-profit institutions	
GDP Gross fixed capital formation by households in dwellings	
GDP Gross fixed capital formation by public administrations	

GDP Changes in stocks + Acquisitions less disposals of valuables	
GDP Changes in inventories	
GDP Acquisitions less disposals of valuables	
GDP External balance of goods and services	
GDP Exports of goods and services	
GDP Imports of goods and services	
Population (in thousand persons)	
Weather variables	
Sun spots	Sunspots are temporary phenomena on the Sun's photosphere that appear as spots darker (cooler) than the surrounding areas.
Degree days	Degree days is a measurement designed to quantify the demand for energy needed to heat a building.

1.5. ANTICIPATED OUTCOMES OF THE STUDY

One of the desirable outcomes of the prediction exercise is an improvement in the process of reporting. The new process would run alongside the current reporting process or will be a part of it. In this way it does not undermine the current process, and hence is without risk. In an ideal situation, the prediction tasks could replace the existing process and we could use the prediction as an official forecast.

VITO has already outlined a broader strategy to research the use of ML. However, the department where the energy balance report is produced is not currently leading any other initiatives to research the use of Machine Learning. The second potential outcome of the study is therefore the opportunity to use the results of this study in other areas of work at VITO. Techniques to solve the missing data problem can be shared and applied in other projects. We can then pass on the lessons learned and best practices when using machine learning.

CHAPTER 2 DATA

2.1. DATA SETS

All data used in this study is quarterly data, ranging from Q1 2000 through Q1 2019. We distinguish target variables from auxiliary variables. The target variables are the variables of interest, in this case the energy variables, see Table1. These are plotted in Fig 1.

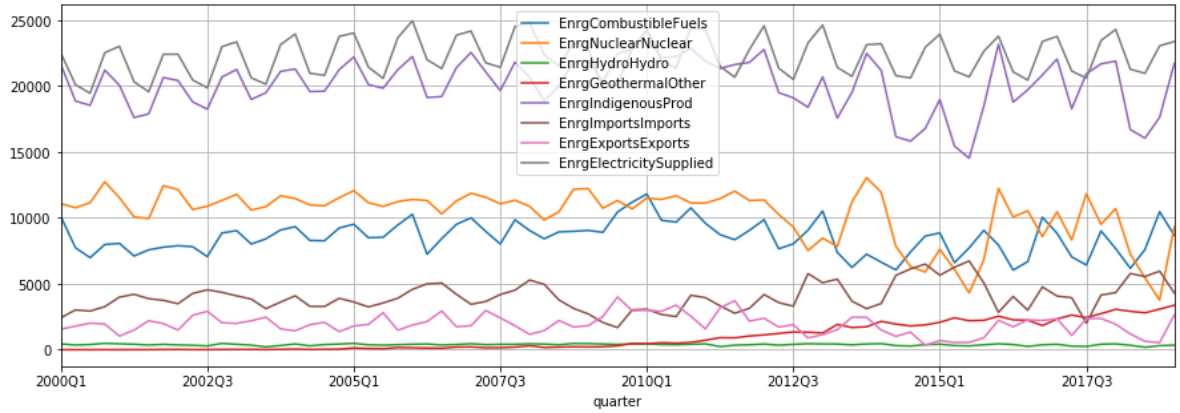

Fig. 1: Series of the 8 energy variables under consideration.

Auxiliary variables – used as predictors – can be any conceivable series, available for the same time period and at the same frequency. In this study economic and weather indicators are considered, see Table 1.

In addition to these variables, we consider past observations of the energy variables as auxiliary data too. Specifically, we include for each energy variable one previous observation, in particular that of 4 time periods earlier, which is the value of the variable in the same quarter one year earlier.

2.2. INPUT DATA

We split the data into two sets, one for training the models, and one for testing. We use about 80% of the data for training: 63 time periods, from Q1 2000 up to and including Q1 2015. The remaining 16 quarters, from Q2 2015 through Q1 2019 are used for testing.

The input to the models, in the training phase, are the auxiliary variables in the training set. For training, the target variables are also used by the algorithms to optimize the model fit.

For testing, the target variables in the test set are not used by the Machine Learning algorithms. In this case, the auxiliary data of the test set is simply used as input to the trained algorithm. The target variables in the test set are used to assess predictive performance of the models.

2.3. DATA PREPARATION

Some algorithms can become numerically unstable when variables have widely varying ranges and values. Therefore the data is standardized before applying algorithms, by computing the standard score z of a sample x as $z = (x - u) / s$, with u the mean and s the standard deviation of the samples in the training set. The test set is scaled in the same manner, but is not used to compute u and s. Such transformation is applied to both the auxiliary data as well as the energy data.

2.4. FEATURE SELECTION

The advantage of many machine learning algorithms is that they can cope with high dimensional auxiliary data. For such algorithms features need not be selected explicitly, rather, the algorithms will decide about their importance implicitly.

Of the methods considered in this study, linear regression models – which can hardly be called machine learning – are an exception. Considering the size of our data set we decided to retain for each regression model the top three features, which we determined using *recursive feature elimination* (see 3.2.1 for details).

2.5. OUTPUT DATA

The output of the models are predictions for the energy variables based on the input data. Using the trained model, predictions are made for the energy variables for the period Q2 2015 – Q1 2019, and compared with the known, true, values.

Since the data were normalized, the outputs of the models are back-transformed to the original scale before assessing predictive accuracy and quality of the predictions.

CHAPTER 3 MACHINE LEARNING

3.1. BASELINE ESTIMATES

In order to compare more advanced models and algorithms, we compute a baseline estimate that will be used as a benchmark. This baseline estimate is very easy and simple to compute, and basically predicts a value of the series *y* at time *t* with the value of *y* one year earlier

$$y_t = y_{t-4}$$

where the *-4* refers to 4 time quarters earlier, as in our setting time periods are quarters. No uncertainty measure is computed for this estimate.
As mentioned in section 2.1, these past values y_{t-4} are included in the set of predictors used in the non-benchmark models. Hence, these other models can be regarded as extensions of the benchmark both in terms of data they use as well as in terms of prediction models or algorithms.

3.2. MODELS AND ALGORITHMS

3.2.1. LINEAR REGRESSION MODEL

In a linear regression modeling approach, all auxiliary variables are considered as independent variables, and the energy variable as the dependent variable. Since there is a limited number of data samples in the training set – only 61 samples – we restricted the number of independent variables in the model to 3. Recursive feature elimination (RFE) is used. In RFE, we start with the full model. Which predictors are retained is decided based on the predictive accuracy of the model that remains after eliminating one predictor. The process is repeated until the desired number of predictors are left over, in this case three. We are concerned that linear regression models with more predictors would run a risk of over fitting and rather choose fairly parsimonious models. Gelman & Hill (2007) is one of many reference texts on regression models.

3.2.2. RIDGE AND LASSO REGRESSION

Ridge and lasso regression are two types of penalty methods, well covered by Hastie et al. (2009), for example. Where ordinary least squares regression aims to minimize the mean squared error, penalty methods include a penalty term which is small for numerically small regression coefficients, favoring those as opposed to numerically large ones. This stabilizes the regression, makes it more robust and avoids over fitting, without the need to select a subset of predictors. Lasso regression uses the L1 norm (absolute values) while ridge regression uses the L2 norm (squared values) in the penalty term.

3.2.3. RANDOM FOREST

A random forest regressor is an ensemble of many regression trees (Breiman, 2001). One tree is an algorithm that splits a data set into smaller and smaller groups, minimizing the within variance and maximizing the between variance. The idea is that samples with similar target variable values get

sorted together based on their auxiliary characteristics. When constructing trees, one can choose the depth of the tree, as well as the required minimum samples at leaf nodes. The random forest technique applies many trees, each time with only a subset of the available predictors. Random forests are intended to be more robust than single trees. No variable selection is needed, the algorithm will choose the important ones automatically.

3.2.4. NEURAL NETWORK

Neural networks are loosely based on ideas of how the brain works, and are essentially algorithms composed of compute units called nodes, that take multiple inputs and compute a single output by applying a function – called activation function – to the inputs. The output from one node can serve as the input of another node. Each time a value is passed on, it gets multiplied by a weight; these weights are adaptable and are learned from the data. There are many possible architectures of neural networks. A common one is the multilayer perceptron, a feedforward network in which several layers consisting of multiple nodes each are used (Hastie et al, 2009). Neural networks are flexible and mostly advantageous to use in cases where the relations between the inputs (predictors) and the output (target variable) are non-linear.

3.2.5. ENSEMBLE PREDICTION

When no single model performs best at all times for all variables, one could apply model averaging and produce ensemble predictions, combining multiple predictions from other models. In this study we produce a simple ensemble prediction by computing the mean of the five predictors: linear regression, ridge regression, lasso regression, random forest and neural network. For now these five estimates are weighted equally. A more advanced ensemble approach would be to weight the underlying estimates with the inverse of their variances.

3.3. MODEL EVALUATION AND SELECTION

The approaches listed in the previous section are evaluated by assessing the predictive accuracy on the test set. The test set consists of 4 years of data: n = 16 data points corresponding to quarters. The predictions $\hat{y}_i$ are compared with the true values y_i using the following measures of predictive accuracy

Root mean squared error

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{y}_i - y_i)^2}$$

Mean error

$$ME = \frac{1}{n}\sum(\hat{y}_i - y_i)$$

Mean absolute error

$$MAE = \frac{1}{n}\sum|\hat{y}_i - y_i|$$

Mean absolute percentage error

$$MAPE = \frac{1}{n}\sum\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

3.4. HARDWARE AND PERFORMANCE

The data set and methods are not of that nature that dedicated high performance computing infrastructure is needed. The implementation has been conducted on a standard Windows 10 computer, with Intel i7 CPU with 6 cores, and 32 GB of RAM. Conducting the complete analysis reported in this study takes approximately 4 minutes. Note that we did not perform cross validation, bootstrapping or other resampling approaches – these would increase the runtime considerably.

All analyses are programmed in Python and are publicly available, see section 5.1.

CHAPTER 4 RESULTS

There are 8 energy related variables that we studied in this project. We applied the methods described in Chapter 3 to the data divided into training and test sets as discussed in Chapter 2.

For each of the target variables the four quality measures RMSE, ME, MAE and MAPE were computed. The results are given in Table 2.

The random forest and neural network methods are intrinsically stochastic. Therefore we fixed a random seed in the code for these methods, so that the results in Table 2 can be reproduced. See section 5.1 on availability of python code.

Table 2. Results of the study. The cells with the best scores are highlighted for each variable.

Variable	methods	RMSE	MAE	ME	MAPE
EnrgCombustibleFuels	yminus4	913	853	-123	10,8%
	linmod	1251	1064	88	13,5%
	ridge	1492	1200	-690	14,6%
	lasso	1570	1408	-402	18,2%
	randomforest	1188	1039	544	14,1%
	neural net	2022	1542	-985	18,5%
	ensemble	1321	1114	-289	13,8%
EnrgElectricitySupplied	yminus4	382	287	-25	1,3%
	linmod	750	655	-416	3,0%
	ridge	1129	941	889	4,2%
	lasso	1185	1048	941	4,7%
	randomforest	425	340	253	1,5%
	neural net	438	367	-347	1,6%
	ensemble	466	378	264	1,7%
EnrgExportsExports	yminus4	1032	834	-101	81,0%
	linmod	921	761	335	91,1%
	ridge	1299	1065	-784	69,7%
	lasso	1318	1061	-743	74,1%
	randomforest	888	794	-163	70,5%
	neural net	1162	914	-626	58,1%
	ensemble	998	842	-396	62,9%
EnrgGeothermalOther	yminus4	425	373	-279	14,5%
	linmod	2105	2075	-2075	80,6%
	ridge	453	391	-310	15,0%
	lasso	495	417	-348	15,6%
	randomforest	812	712	-706	26,0%
	neural net	1182	1116	-1116	42,1%
	ensemble	975	911	-911	34,2%

EnrgHydroHydro	yminus4	61	47	14	14,9%
	linmod	80	64	49	23,2%
	ridge	66	51	23	18,3%
	lasso	77	60	53	22,5%
	randomforest	65	49	37	18,2%
	neural net	122	104	104	36,9%
	ensemble	79	62	53	22,8%
EnrgImportsImports	yminus4	1797	1382	145	35,0%
	linmod	1420	1120	-638	25,3%
	ridge	1882	1606	1239	44,6%
	lasso	2098	1804	1488	50,3%
	randomforest	1626	1377	953	39,7%
	neural net	1376	1146	446	32,3%
	ensemble	1479	1223	698	34,9%
EnrgIndigenousProd	yminus4	2658	2084	-271	11,1%
	linmod	3183	2596	1674	14,9%
	ridge	2595	2138	-1420	10,7%
	lasso	2773	2222	-1625	11,1%
	randomforest	2254	1944	-654	10,2%
	neural net	2509	2208	-1403	11,1%
	ensemble	2222	1947	-686	10,1%
EnrgNuclearNuclear	yminus4	3284	2614	117	37,8%
	linmod	3509	2761	2495	47,5%
	ridge	2902	2370	-581	33,3%
	lasso	2756	2300	-435	32,2%
	randomforest	2681	2369	-902	31,6%
	neural net	2847	2389	-11	34,1%
	ensemble	2580	2145	113	32,4%

For four of the eight variables, the *yminus4* baseline method performs best according to almost all criteria. The relatively good performance can have different explanations. For the variable *EnrgElectricitySupplied*, the baseline predictions are very good already, with a MAPE of only 1,3%, a score that is hard to beat. The variables *EnrgCombustibleFuels*, *EnrgGeothermalOther* and *EnrgHydroHydro* have the best scores for the *yminus4* method as well, but their MAPE is somewhat larger. The ML methods are not able to improve upon the baseline method. The most likely explanation is that the predictors that are used in this study do not have enough predictive power.

For *EnrgIndigenousProd*, the MAPE for *yminus4* is not very large either, nevertheless, the ML methods improve predictions. In this case *randomforest* performs rather well. When it comes to MAPE, the *ensemble* approach scores best. It is known but remains astonishing that a method that is the average of several other methods scores better than each of these methods individually.

Results from the different criteria not always point to the same method. For *EnrgExportsExports*, random forest scores best on RMSE, the linear model is best on MAE, the baseline on ME and the neural network on MAPE.

For *EnrgNuclearNuclear*, the *ensemble* method has the best RMSE and MAE scores, while the best MAPE is achieved with the *random forest*, and no method beats the neural network when considering ME.

The *linmod* method performs well for *EnrgImportsImports*. The ridge and lasso regressions never come out as the best method in our application. However, whenever the baseline method is beaten by an ML method, they tend to beat it as well, although never to the largest extent.

In the following table, we ranked the 7 methods within the 8 predicted variables from 1 to 7 for each quality measure. A low rank indicates a high quality measure value (bad predictions). The higher the rank the lower the measure, in this way a high rank is better (good predictions). For the Measurement Error (ME) measure we ranked ABS(ME) – its absolute value – because values closer to zero are preferred.
We see that Random Forest has the highest rank on 3 quality measures (MAE, MAPE and RMSE). Yminus4 has the best score for the quality measure ABS(ME) and has the best grand total.
We choose to show the results for all the quality measures on the grounds that all quality measures have their merits, and are commonly used in model evaluation studies. Of the four measures shown here, our personal view is that if a single measure were to be used, it should be MAPE, since it expresses errors percentage-wise.

Quality measure	**Average of yminus4**	**Average of linmod**	**Average of ridge**	**Average of lasso**	**Average of randomforest**	**Average of neural net**	**Average of ensemble**
ABS(ME)	6,6	3,4	3,1	2,6	4,4	3,5	4,4
MAE	5,4	3,4	3,4	2,9	5,5	3,0	4,5
MAPE	4,8	2,8	4,0	3,0	5,0	3,4	4,8
RMSE	5,0	3,1	3,4	2,6	5,6	3,5	4,8
Grand Total	**5,4**	**3,2**	**3,5**	**2,8**	**5,1**	**3,3**	**4,6**

CHAPTER 5 DISCUSSION

5.1. REPRODUCIBILITY

The data sets that are used in this study are all publicly available. For the purpose of this study they are combined into a single data set. To the best of our knowledge and after carefully checking data access rights, the data can be redistributed. We put the data set used in this study online with Zenodo, an open research data and software platform. The data set is available directly from Zenodo at:
https://zenodo.org/record/3596695
However, access is recommended through its DOI, which resolves to the version used in the present report, at the following URL:
https://doi.org/10.5281/zenodo.3596694

The code by means of which the analyses in this study was implemented is being made available as a set of Jupyter notebooks containing Python code. These can be found at the following URL:
https://github.com/VITObelgium/energy-balance-ml

5.2. EVOLUTION OF THIS STUDY INSIDE THE ORGANIZATION

The demand for earlier energy estimates had been recognized within our organization for some time. The opportunity to join the UNECE Machine Learning project in 2019 stimulated us to take this up and to undertake the research that we had wanted to do. During execution of the project, two elements started to play an important role regarding the enthusiasm and support from management and colleagues within the organization.

The first element is that initial results indicated that the gains to be had from ML approaches are not overwhelming, and do not always improve on rather simple baseline methods. ML methods are perceived less transparent while not always delivering impressive results. Nevertheless, we show in this report that ML methods have promise and that there are certain variables for which early estimates could be improved using ML. In our case, applying ML does not give quick wins, probably to a large extent because of the weak predictive power of the predictors under study. Sustained support and substantial research efforts are needed to make this research a success. It is the view of the authors that we may need to manage expectations better, in times when the media present success after success of AI and ML. In reality, success comes at a cost, not quickly and easily.

A second element is organizational, not technical. VITO will discontinue the production of the energy balances from 2021 onwards. The activities will be taken over by the Flemish Energy Agency who are already directly involved in energy data collection at the most detailed level. VITO will continue to conduct research on energy related issues in a vast range of research projects and contracts for government and industry and may still benefit from the results of this study, be it in another setting than originally intended.

5.3. STATUS OF THIS PROJECT

The current state of affairs of this project is reflected in this document. Considering the arguments mentioned in the previous section, it is unlikely that significant effort will be committed to further this research in its current form. For a continuation to happen, the research goals should be revisited. In addition, it would be greatly beneficial to find a source of funding for such research, as without explicit funding no substantial investment in efforts and resources will be made.

5.4. RISKS

The research reported here is without risk in the sense that this project never aimed at changing the current production process of the energy balance statistics. That process can continue to exist as it always has. However, as outlined in section 5.2, when the responsibility of producing these statistics is transferred to another agency, they may want to choose to implement changes to the process as they see fit. The success or not of our research reported here has no impact on these events.

5.5. COLLABORATION WITH OTHER INSTITUTES

VITO did not collaborate with other institutes in the context of the research reported here. However, other participants of the UNECE Machine Learning project expressed their interest in our results and our data. Given that besides this report, also the data and code are publicly available (see section 5.1) it might be expected that at some point our work will be complemented by others.

CHAPTER 6 CONCLUSIONS

6.1. LESSONS LEARNED

We summarize our lessons learned in a bullet point fashion and distinguish between technical and organizational aspects.

Technical:

- expect to spend a considerable amount of time on collecting and preparing data sets,
- think of a baseline method that is simple, common-sensical and reasonably performing; this is to avoid drowning in complexities with only marginal effects,
- no single ML method worked best for us; when ML methods improved on the baseline method, random forests and neural networks worked well, and linear regression to a lesser extent; ridge and lasso regressions seemed not useful in our setting,
- in our study the ensemble method, averaging results from several ML methods, seems promising.

Organizational:

- manage expectations well; some people expect great results without effort or investment; low hanging fruit is sparse and there are no magic bullets,
- substantial effort is needed to conduct a proper investigation into the usability of ML methods,
- it is best to find funding and dedicated time for such a project to allow for continued progress, resources and capacity,
- it is important to report on findings and not let results be lost,
- making data and code publicly available has been well received by the community and can stimulate future joint work.

6.2. NEXT STEPS

There are a number of actions that could be undertaken to further the research reported on here. However, it is uncertain whether the authors will take these up in the short term.

With respect to the ML methods applied in this study, tuning of the hyper parameters could be conducted. Now, we have based these to some extent on ad hoc trial-and-error experiments. For example in the neural networks, the choice of the number of layers and nodes, or in random forests the number of trees and the construction of the trees. With proper hyperparameter tuning the performance of some of the methods might improve to some degree. In addition, there are popular machine learning methods that we did not apply yet, such as support vector machines.

In this study the temporal nature of the data has not been fully exploited. Other than including one past observation of the energy variables, we didn't account for time dependencies in any of the models. From the ARIMA-family of time series models we basically used an AR(4) model with

additional regressors as our linear model, but did not investigate integration or moving average regression. The machine learning models we used are not aimed at time series data specifically. An important reason is that the series we have are quite short in machine learning terms – nevertheless, a closer investigation might be worthwhile. Finally, because of the clear seasonal patterns, structural time series models with seasonal components could prove useful.

An important aspect we have not addressed so far is quantification of uncertainty. This is an area receiving much attention lately, and concerns the estimation of the uncertainty associated with point estimates. Bootstrap procedures are a popular and common approach that could be implemented in our project.

In addition to the electricity variables considered in this study, more variables from the energy balance report could be analyzed, for example supply of natural gas. Other potential extensions of this work include disaggregating the data by industry sector.

It may be expected that within the UNECE Machine Learning project advice and recommendations will be circulated with the suggestion for implementation in the case studies such as our project. Such suggestions might come from Work Package 2 on quality aspects. It remains to be seen what the suggestions entail exactly and if and how they can be applied in our study.

REFERENCES

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Claeskens, G. & Hjort, N. L. (2008). Model Selection and Model Averaging. Cambridge: Cambridge University Press.

Daas, P.J.H., Puts, M.J., Buelens, B. and van den Hurk, P. (2015). Big data as a source for official statistics. Journal of Official Statistics, 31, 249–262.

Gelman, A. & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models, Vol. 1 New York: Cambridge University Press.

Hassani, H., Saporta, G. and Silva, E.S. (2014). Data mining and official statistics: the past, the present and the future. Big Data, 1, 34–43.

Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed. New York: Springer.

UNECE Machine Learning Team (2018). The use of machine learning in official statistics. Position paper, available at https://statswiki.unece.org/download/attachments/261818141/The%20use%20of%20machine%20learning%20in%20official%20statistics.pdf