TÉCNICO LISBOA

Automatic classification of LCA data

Shanlin Chen

Thesis to obtain the Master of Science Degree in

Energy Engineering and Management

Supervisors: Dr. Christopher Lucien Mutel

Prof. Susana Margarida da Silva Vieira

Examination Committee

Chairperson: Prof. Edgar Caetano Fernandes

Supervisor: Prof. Susana Margarida da Silva Vieira

Member of the Committee: Prof. Carlos Augusto Santos Silva

June 2020

Acknowledgement

I would like to express my gratitude to all the persons who made contributions to the work described in the thesis.

First I would like to thank Dr. Christopher Lucien Mutel and Prof. Susana Margarida da Silva Vieira for their great supervision and support during my thesis work, their experience, knowledge and regular inputs helped to move my steps forward. This thesis could not be finished without their patient guidance.

I would also like to thank Dr. Romain Sacchi for providing information and remarkable insights to bring machine learning into Life Cycle Assessment studies.

Next I wish to thank PAUL SCHERRER INSTITUT for the opportunity to write my thesis and all the materials, resources and thoughtful supports I received.

Finally, I would like to thank InnoEnergy, SUT and IST that allowed me to follow my studies in an incurably inspiring environment.

Abstract

Life Cycle Assessment (LCA) is a useful tool for environment impact assessment and decision making in energy systems, meanwhile it's also complicated as it requires detailed information, time and computing resources. Many studies have been conducted to make LCA simpler in the categories of impact assessment like Cumulative Fossil Energy Demand, Carbon Footprints and Representativeness Index. Life Cycle Inventory (LCI), as an important part of a LCA study compiles all necessary information. This study aims to use machine learning algorithms to explore the data structures and hidden patterns in ECOINVENT and EXIOBASE LCI database, estimate the LCI of wind turbines with limited information for LCA.

Clustering results from different machine learning algorithms show that all the datasets of ECOINVENT or EXIOBASE seems to be one cluster with some random outliers by using inventories as features. Location is a good feature to change the data structure but more information needed for better clustering.

It's possible to estimate the LCI of a wind turbine from partial information through machine learning algorithms and mathematical methods. It's also practical to predict the total electricity production of wind turbines for better environmental impact assessment. Although the estimations of input materials and predicted amount of electricity may differ from the actual values, they are still good references for impact assessment as well as decision making.

Key-words: Life Cycle Assessment, Life Cycle Inventory, Machine Learning, Wind Turbine

Resumo

A Avaliação do Ciclo de Vida (ACV) é uma ferramenta útil para avaliação de impacto ambiental e tomada de decisão em sistemas de energia, mas também é complicado, pois requer informações detalhadas, tempo e recursos de computação. Muitos estudos foram realizados para tornar a ACV mais simples nas categorias de avaliação de impacto, como Demanda Cumulativa de Energia Fóssil, Pegadas de Carbono e Índice de Representatividade. O Inventário do Ciclo de Vida (ICV), como parte importante de um estudo de ACV, compila todas as informações necessárias. Este estudo tem como objetivo usar algoritmos de *machine learning* para explorar as estruturas de dados e padrões ocultos no banco de dados ICV ECOINVENT e EXIOBASE, e estimar o ICV de turbinas eólicas com informações limitadas para ACV.

Os resultados de *cluster* de diferentes algoritmos de *machine learning* mostram que todos os conjuntos de dados de ECOINVENT ou EXIOBASE parecem ser um cluster com alguns outliers aleatórios usando inventários como recursos; a localização é um bom recurso para alterar a estrutura de dados, mas outras informações são necessárias para um melhor cluster.

É possível prever o ICV de uma turbina eólica a partir de informações parciais por meio de *machine learning* e métodos matemáticos, também é viável prever a produção total de eletricidade de turbinas eólicas para melhor avaliação do impacto ambiental. Embora as estimativas de matérias primas e quantidade prevista de eletricidade possam diferir dos valores reais, elas ainda são boas referências para avaliação de impacto e também de tomada de decisão.

Palavras-chave: Avaliação do Ciclo de Vida, Inventário do Ciclo de Vida, *Machine learning*, Turbina eólica

Content

Figures

Tables

Symbols

Symbol	Meaning
A	technosphere matrix
a	mean distance between a sample and all other points in the same class
B	biosphere matrix
b	mean distance between a sample and all other points in the next nearest cluster
C	characterization matrix
Cf	capacity factor
f	final demand vector
h	characterized inventory matrix
i	ith number in the dataset
max	maximum value
n	number of samples
p	number of products
P	rated power
q	number of activities
r	number of elementary flows
R	rotor diameter
R^2	coefficient of determination of the prediction
S	silhouette coefficient
u	residual sum of squares
v	sum of squares of difference between true and mean values
$w(w_1, w_2, \ldots, w_p)$	coefficients
$x(x_1, x_2, \ldots, x_p)$	input variables
$y(y_1, y_2, \ldots, y_p)$	target values
y_{pred}	predicted values
y_{true}	actual values
$\overline{y_{true}}$	average of y_{true}

Abbreviations

Abbreviation	Meaning
ALOP	Agricultural Land Occupation
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CF	Capacity Factor
CPC	Central Product Classification
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FA	Factor Analysis
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HDPE	High-density Polyethylene
ISIC	International Standard Industrial Classification
LCA	Life Cycle Assessment
LCI	Life Cycle Inventory
LCIA	Life Cycle Impact Assessment
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MR-IOT	Multi-Regional Input-Output
MR-SUT	Multi-Regional Environmentally Extended Supply-Use
MSE	Mean Squared Error
Multi-LCA	Multiple Life Cycle Assessment
NN	Neural Network
OPTICS	Ordering Points to Identify the Clustering Structure
PC	Principle Component
PCA	Principle Component Analysis
PP	Polypropylene
PVC	Polyvinyl Chloride
RES	Renewable Energy Systems
RF	Random Forest
RMSE	Root of Mean Squared Error
SVM	Support Vector Machine
WTG	Wind Turbine Generator

Chapter 1

Introduction

This chapter gives some background information about Life Cycle Assessment (LCA), Life Cycle Inventory (LCI) databases like ECOINVENT and EXIOBASE, as well as some related studies trying to make LCA simpler with regards to impact categories. Then the main purposes of this work, and some introduction about Brightway2, which is an open source software for LCA calculations are also provided.

1.1 Life Cycle Assessment

Life cycle assessment (LCA) is a technique compiling an inventory of relevant inputs and outputs of a product system, evaluating the potential environmental impacts associated with those inputs and outputs, and interpreting the results of the inventory and impact phases in relation to the objectives of the study. An LCA study consists of a thorough inventory of energy and materials that are consumed during the whole process related to the product, process or service, and identify and evaluate the corresponding emissions to the environment. Mainly there are four steps in a LCA study as shown in Figure 1, goal and scope definition, inventory analysis, impact assessment and interpretation.

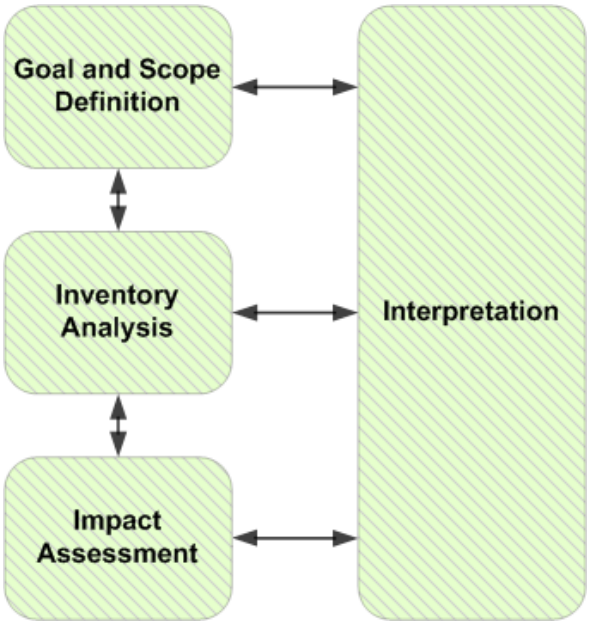

Figure 1-Illustration of the general phases of LCA, as defined by ISO 14040

Goal and scope definition gives the objectives and system boundary. After definition of the goal and scope, inventory analysis is the LCA step involves the compilation and quantification of inputs and outputs for a given product system throughout its life cycle in the system boundary. Life cycle inventory (LCI) includes the collected data and its compilation from inventory analysis.

Impact assessment is trying to identify and evaluate the amount and importance of the potential environmental impacts resulting from the LCI based on different LCIA methods. The inputs and outputs will be assigned to impact categories and their potential environmental impacts are quantified from characterization factors.

Then the results from impact assessment can be utilized to evaluate systems for improvements or

decision-making with regards to the environment and sustainability.

LCA is a very complex tool that requires a large amount of detailed information and also time. Many studies have been conducted to offer possibilities to make LCA easier for implementation and interpretation but more on the basis of impact categories. Cumulative fossil energy demand might be a useful indicator for environment performance, because combustion of fossil fuels is a major contributor to many environmental problems, it could be a screen indicator for energy production, material production, transport, global warming and resource depletion, but not for waste treatment and land use [1]. Representativeness index aims to reduce the size of environment impact categories by correlation analysis, which also contributes to the interpretation of LCA results through pointing to specificities of inventories and identifying the main representative impact categories [2]. Resource footprint and damage footprint may be good proxies of environmental damage, while resource footprints accounted for more than 90% of the variation in the damage footprints [3]. And studies aimed at comparing different LCIA methods generally find they are highly correlated [3,4], so it's not very influential to choose the impact assessment methods and using representatives for other LCIA methods can simplify LCA with less calculation based on the strong correlations between the impact categories.

LCI is compiled of all the necessary information for LCA studies, which can be provided by LCI databases like ECOINVENT [5] and EXIOBASE [6]. The ECOINVENT is a widely used LCI database offers full information for over 18000 datasets and 3000 products in up to 140 countries. ECOINVENT datasets cover all relevant environmental flows, including resource extraction, land use, emissions as well as materials and energy supplies in a global or regional level. EXIOBASE consists of global Multi-Regional Environmentally Extended Supply-Use (MR-SUT) table and Input-Output (MR-IOT) table. A large number of countries are involved via MR-SUT and MR-IOTs for estimating emissions and resource extractions by industry, as well as analysis of environmental impacts related with the final consumption of products.

1.2 Objectives

There is a lot of detailed information in LCI databases, like electricity production and transport. They are split into groups based on engineering judgments. However, we can use machine learning algorithms to create groups based on their measured attributes, including LCI, LCA indicators. Using data-based classification could reduce uncertainty and better help us understand when to or not to split datasets

into different groups, which should in turn help us understand the data better.

Exploring the data patterns in LCI databases is the first part of this thesis work, which aims to understand and use data in a more efficient way. Applying machine learning algorithms in creating LCI of wind turbines is the second part, which gives a possibility to perform LCA with even limited information.

Exploring the hidden information in LCI databases is trying to apply unsupervised machine learning methods to discover the patterns of unlabeled data, two datasets from ECOINVENT and EXIOBASE will be used, and the input features can be their inventories or LCA scores from different LCIA methods, meanwhile, different techniques for dimensionality reduction and different clustering algorithms will be applied and compared.

LCI estimation aims to apply supervised machine learning algorithms to estimate the detailed LCI of wind turbines for LCA studies by using limited information like capacity, dimensional size and the location (onshore or offshore). A wind turbine is usually made up of rotor, nacelle, tower, transformer, foundation and some power cables as shown in Figure 2, which is an onshore wind turbine, but for offshore wind turbines, the structure is similar apart from the foundation, there are many different types for offshore foundations, and monopile is the one assumed for offshore wind turbines in this work, as it takes up to 80% of the all offshore foundations.

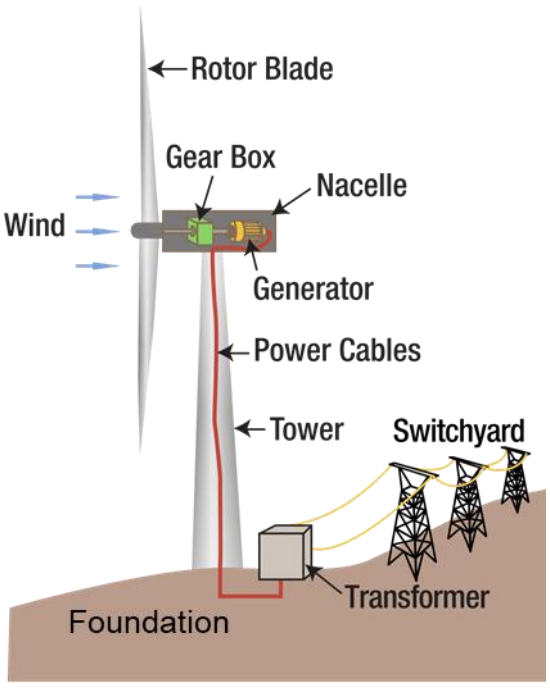

Figure 2-The structure of an onshore wind turbine

There are available datasets for training machine learning models for sizing rotor, nacelle and tower, for the other parts, some mathematic models are applied to size foundation, transformers and power cables.

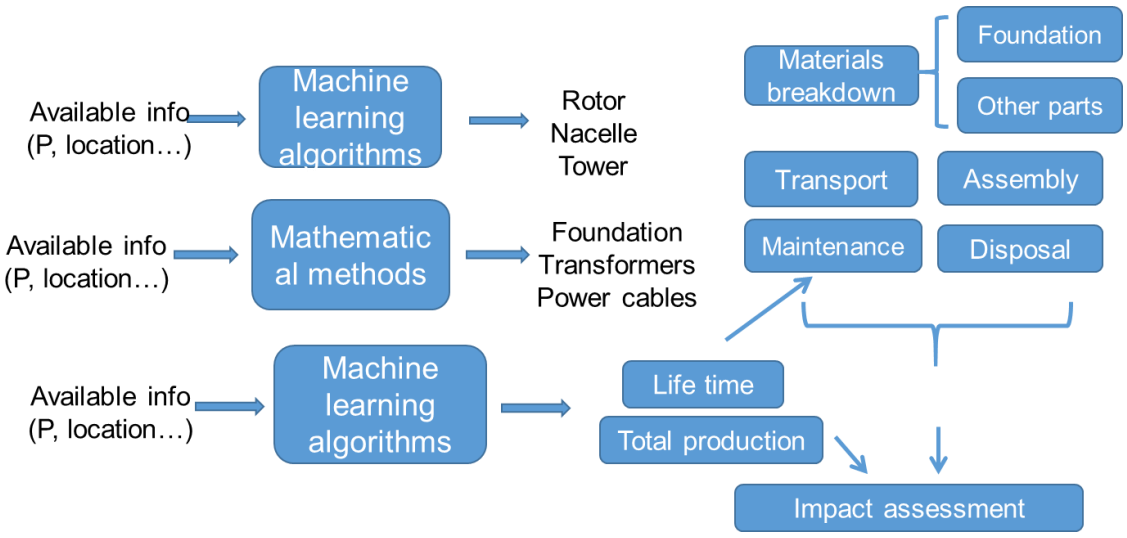

Figure 3-The path for LCI estimation and impact assessment

Figure 3 shows a path for LCI estimation and impact assessment. After sizing all the components of the wind turbine, the materials breakdown can be achieved by different strategies based on the rated power, and the activities involved in transport, assembly and end_of_life disposal are based on the input materials and some assumptions, and maintenance is determined by life time and size of the wind farm. With detailed LCI estimation and the total electricity production predicted by the machine learning model, it's now possible to perform the impact assessment by Brightway2.

1.3 Brightway2

LCA calculations in this study are performed by Brightway2 [7], which is an open source framework for LCA, it's designed to be easy to use, especially when used with Jupyter notebooks [8]. In Brightway2, LCA can be expressed in a single formula:

$$h = CBA^{-1}f \quad (1)$$

Where:

- *h* is the characterized inventory matrix (dimension *1 x 1*), which can be seen as the LCA score from a certain LCIA method.

- *C* is the characterization matrix (dimension *r x r*), which is defined by different LCIA method. In Brightway2, each impact assessment method is a set of characterization factors for a set of biosphere flows. Each impact category and subcategory is a separate method, and each method is stored and calculated separately.
- *B* is the biosphere matrix (dimension *r x q*), which describes the exchanges between the environment and the activities. Biosphere database has all the resource and emission flows from the ECOINVENT database.
- *A* is the technosphere matrix (dimension *p x q*), which defines the inputs for different activities, like energy, materials requirements.
- *f* is the final demand vector (dimension *p x 1*), which is total demand during the whole process of a product or a service.

And the numbers for dimensions:

- *p* is the number of products
- *q* is the number of activities
- *r* is the number of elementary flows (bio flows)

Based on the formula and specific parts also have different names:

- $A^{-1}f$ is called supply array
- $BA^{-1}f$ is the inventory

Brightway2 also involves Multi-LCA calculations with a set of activities and LCIA methods, which can be used in LCA of wind turbines through LCI.

1.4 Contributions and outlines

Although it is a commonsense that wind turbines with higher capacities will have larger size of the components, it doesn't mean there is a linear relationship. Different machine learning algorithms like Linear Regression, Random Forest, Neural Network and Support Vector Machine are applied to train the models, and comparisons are also made with these algorithms.

Meanwhile, there is also a comparison of LCI estimation of wind turbines with some actual values from some LCA reports, reasons for the differences between the input materials are explained. Regarding the impact assessment, the calculated results from estimated LCI and values from LCA reports are compared for global warming potential, the differences between them are also described and

explained.

For future work, there is a possibility to apply machine learning methods for all the LCI estimation of wind turbines. An outlook to apply the same methodology for other renewable energy systems is also explained based on powerful machine learning tools and ECOINVENT LCI database.

The outline of this thesis is summarized as following:

Chapter 1: Introduction of the topic, followed by the objectives and research questions. Some background information regarding Life Cycle Assessment and Brightway2 are also provided.

Chapter 2: This chapter aims to apply machine learning methods to explore explore the data structure or hidden information in LCI databases of ECOINVENT and EXIOBASE. Different input features including inventories and LCA scores from different LCIA methods are used, as well as different dimensionality reduction techniques and clustering algorithms.

Chapter 3: LCI estimation of wind turbines is trying to use machine learning algorithms to estimate the detailed LCI for LCA studies with limited information. Due to the lack of sufficient data, some mathematical models are also applied. The estimated results were compared with the values from some LCA reports, there are some differences, and the reasons are explained.

Chapter 4: This chapter summarizes the conclusions from this work including LCI database clustering and LCI estimation of wind turbines. It also points out the aspects to be developed in future works like estimating the entire LCI by machine learning models if there is enough available data, and the methodology of LCI estimation for wind turbines can also be applied to other RES projects.

Chapter 2

LCI database clustering

LCI database clustering aims to explore the data structure or hidden information in LCI databases of ECOINVENT and EXIOBASE. Different input features including inventories and LCA scores from different LCIA methods are used, as well as different dimensionality reduction techniques and clustering algorithms.

2.1 Current classification systems

In a LCI database like ECOINVENT or EXIOBASE, there are usually thousands of processes/activities grouped by International Standard Industrial Classification (ISIC) or Central Product Classification (CPC) from an engineering point of view.

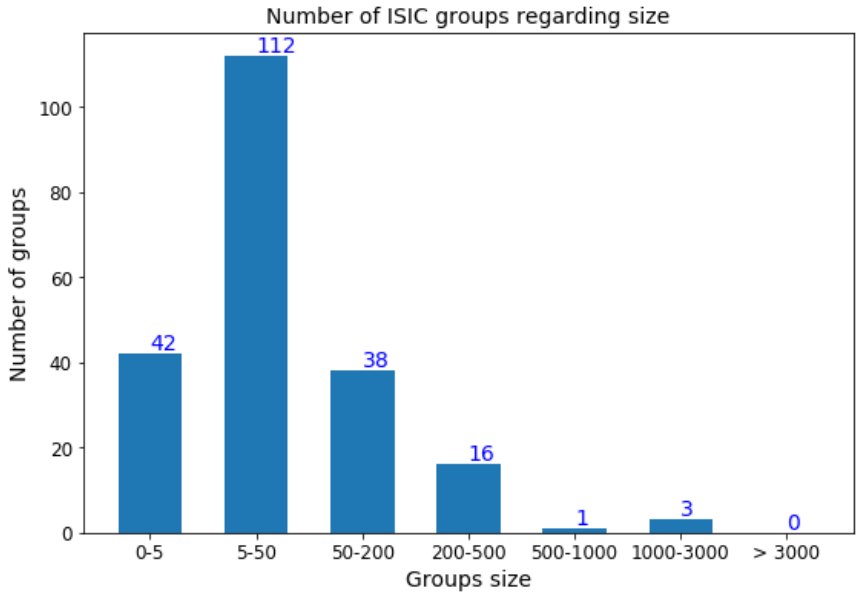

Figure 4-Group size of ISIC classifications for ECOINVENT

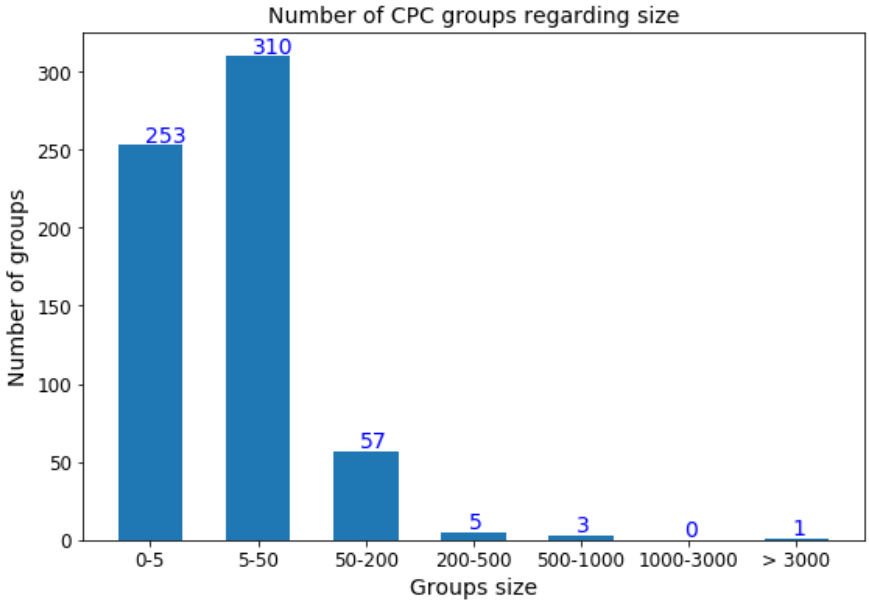

Figure 5-Group size of CPC classifications for ECOINVENT

ISIC is the international reference classification of productive activities, its main purpose is to provide a set of activity categories that can be utilized for the collection and reporting of statistics according to such activities. And CPC consists of a coherent and consistent classification structure for products based on a set of internationally agreed concepts, definitions, principles and classification rules, it is intended to be an international standard for organizing and analyzing data on industrial production, national accounts, trade, prices and so on. Generally, activities are classified according to their main output in ISIC groups, and CPC group is defined by product name.

In fact, about 200 activities/processes do not belong to any ISIC or CPC groups in ECOINVENT. And as shown in Figure 4 and 5, one group may include over 1000 activities, most groups are in the size from 5 to 50, while a number of groups only have 5 or fewer activities.

Activity clustering aims to find the hidden patterns or structures for activities from LCI database by machine learning, with a purpose of better understanding and more efficient utilization. There will be different algorithms being applied and compared. The expected result from clustering would be some different groups found by the algorithms, and there will also be a comparison between these groups and the ISIC/CPC classes.

2.2 Unsupervised machine learning

In pattern recognition problems, the training data is a set of input vector without any corresponding target values are known as unsupervised learning. Unsupervised machine learning can be used for density estimation to determine the distribution of data within the input space, or for visualization by projecting data from a high-dimensional space down to two or three dimensions, or to discover groups of similar samples within the data, which is also called clustering [9].

There are a lot of different algorithms for clustering by using distance, density or similar measurements between data points to identify patterns.

Partitional clustering are methods used to classify samples within a dataset into groups based on their similarity, like k-means [10] algorithm, in which each cluster is represented by the center or means of data points in the cluster, so the k-means method is sensitive to outliers. And partitional clustering methods require the number of clusters before clustering, however, this is rarely the case [11], because the purpose for pattern recognition problems is to find the hidden information, and the number of clusters is a part of this information to be found.

Density-based algorithms consider that clusters are based on connectivity and density functions, like DBSCAN [12], in which the clusters are condensed and separated by areas of low density. Unlike k-means, the clusters found by DBSCAN can be any shape. There are two important parameters to this method, one is the maximum radius of the neighborhood and the other is minimum number of points within the radius of a point. Therefore, DBSCAN can handle outliers but is sensitive to parameters.

Hierarchical clustering is generally a number of family clustering algorithms which find clusters by merging small clusters into larger ones or splitting large clusters. The dendrogram is a hierarchy of clusters and shows how clusters are related. There is no outlier in hierarchical clustering, one cluster can have only one sample. Agglomerative Clustering is a bottom-up approach, which merges small clusters into larger ones based on different linkage criteria [11].

Meanwhile, there are also other clustering methods, OPTICS [13] has many similarities with DBSCAN, while OPTICS builds a reachability graph that change the maximum radius from a single value to a value range. Birch [14] deals with large datasets by compressing the original data into a set of clustering feature nodes, and then clustering the subset instead of initial dataset to reduce the memory requirement. Mean Shift [15] is a centroid based algorithm, which works by updating centroids to be the mean of points in a given region like k-means, but the number of clusters prior clustering is not required.

2.3 Clustering performance evaluation

Like clustering algorithms, there are also many indexes to evaluate the clustering performance. In most case the ground truth classes are not available in clustering problems, so the metrics based on true labels and predicted labels would not be applicable, the evaluations should be performed using the model itself.

The Silhouette Coefficient [16] for a single sample is given as:

$$S = \frac{b-a}{max(a,b)} \tag{2}$$

Where:

- *a* is the mean distance between a sample and all other points in the same class
- *b* is the mean distance between a sample and all other points in the next nearest cluster

The Silhouette Coefficient for the whole dataset is calculated as the mean of the Silhouette Coefficient for every sample.

In the range of -1.0 to 1.0, a higher Silhouette Coefficient means a model has better defined clusters, while scores around 0 mean clusters are overlapped.

Calinski-Harabasz index [17] is the ratio between the sum of between-clusters dispersion and sum of inter-cluster dispersion for all clusters, it's fast to compute and a higher value relates to a model with better results.

Davies-Bouldin index [18] is defined by the average similarity between clusters, and the similarity compares the distance between clusters with the size of the clusters themselves. A lower score means a better separation between the clusters, and the lowest possible score is 0.0.

It is necessary to note that all these three indexes tend to be higher for convex clusters like density based clusters obtained from DBSCAN than others.

2.4 Data preparation

Activity clustering aims to find the main clusters for activities from LCI database. Taking ECOINVENT as an example, the samples are the activities, and the features are the corresponding inventories. As explained before, the $BA^{-1}f$ calculates the inventory, for different activity the demand array will change, that's how to get the corresponding inventory for every activity in ECOINVENT database.

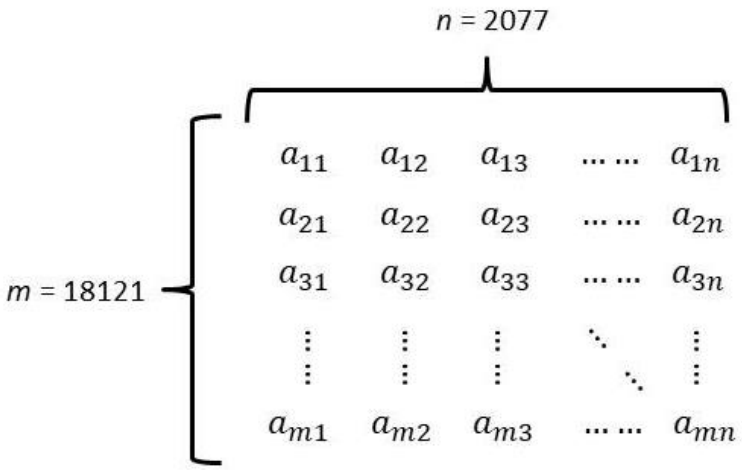

Figure 6-Matrix structure for clustering

The structure of matrix for clustering is shown in Figure 5. The rows are samples which are also activities from ECOINVENT database, the columns are features which are inventories based on elementary flows. The dimensions are 18121 x 2077, which means it's a high dimensional clustering

problem. For activities, there is a dictionary including the activity key and corresponding numbers, and a biosphere dictionary includes the elementary flow code and related numbers, so it is reasonable and easier to replace them by numbers other than names.

Generally, HDBSCAN can handle high dimensional clustering with a dimensionality up to around 50 or 100, performance will decrease significantly beyond that [19]. Subspace clustering is another algorithm working for high dimensional data, which aims to find clusters in different subspaces within a data set [20]. High dimensional data is not only difficult to visualize, but also requires more space and computing power to apply clustering algorithms, that's why we need dimensionality reduction for more efficient handling for large datasets.

Principle Component Analysis (PCA) [21] and Factor Analysis [22] are most frequently used linear techniques for reducing dimensionality. PCA, as a technique for unsupervised dataset, aims to extract important information for a high-dimensional structure into a sub-structure with a lower dimensionality by transforming and reserving essential parts with much more variations. Principal components are the products from PCA, they represent the underneath information of the original data. While Factor Analysis explains the variances among the original features and then condense a lower number of unobserved variables called factors. Each factor explains a curtain amount of information in the initial features. Factor Analysis may also involve PCA for factor extraction and then converts extracted factors into uncorrelated ones.

For selecting the number of dimensions of sub-dataset, there are different strategies. The most traditional and straight method is to plot the explained eigenvalues, which means the variances retained by the principle components, in a descending order as a function of number of dimensions, there will be an elbow point, as shown in Figure 6, after which the slope changes significantly. Another approach is to define how much to retain the percentage of the total variance before the dimensionality reduction, it can be 90% or lower depends on the situations, then a number of components will be generate and give the desired explained variance. In fact, the proportion of the explained variance is calculated from the same way, the main difference is that now the cumulative sum of eigenvalues is the defined value. There are also other ways for dimensionality reduction, like Kernel PCA [23], Isomap [24] for non-linear dimensionality reduction.

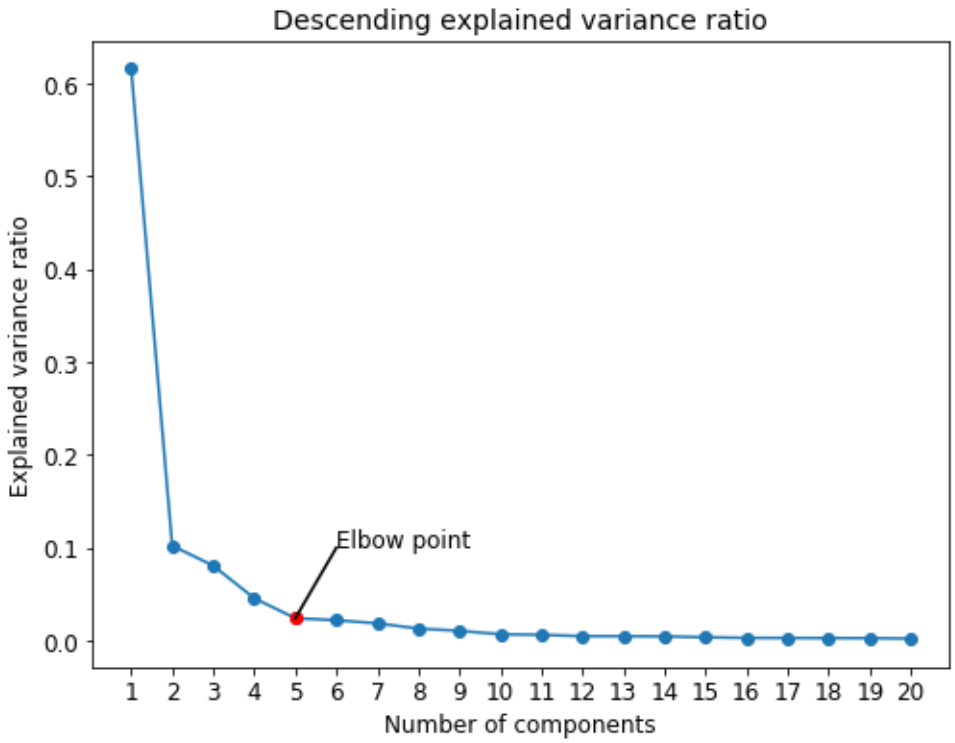

Figure 7-Elbow point to choose the dimension number

It's possible to reduce the dimensionality by engineering solution as well, in this case combining similar elementary flows would work, the elementary flows with the same name and category will be put into one group, for example, Carbon Dioxide Fossil is one group no matter it belongs to 'low population density, long-term' or 'lower stratosphere + upper troposphere'. There will be 1317 biosphere flows remaining after the combination. Correlation analysis among features will find the highly correlated ones (if greater than 0.99) and then the correlated features and zero columns will be removed to reduce the dimensionality further. It is necessary to apply Feature Scaling to standardize or normalize the independent features present in the same range for better performance and less computing power requirement.

2.5 Clustering results

2.5.1 ECOINVENT

Initially there are 2077 bioflows involved in the inventories for an activity, after correlation analysis and PCA, 63 principle components retain more than 90% of the variance. When use these 63 components for clustering, the results from different algorithms are shown in the Table 1. The performance is satisfying as the Silhouette Coefficient is close to 1.0 and the other two indexes are also in the good range, although the Calinski-Harabasz Index of Agglomerative Clustering is far more better than DBSCAN, the results from these two algorithms are quite similar, apart from a small part of outliers, one cluster consists of most samples over 90% as shown in Figure 8 and Figure 9, and the other clusters only have a few samples, which are usually less than 10, this makes the whole dataset more like one cluster with some outliers.

Table 1-Clustering results for ECOINVENT with all bioflows

Clustering algorithm	Number of clusters	Number of noise points	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
DBSCAN	26	407	0.932	96.166	1.946
Agglomerative	25	0	0.973	186478.761	0.424

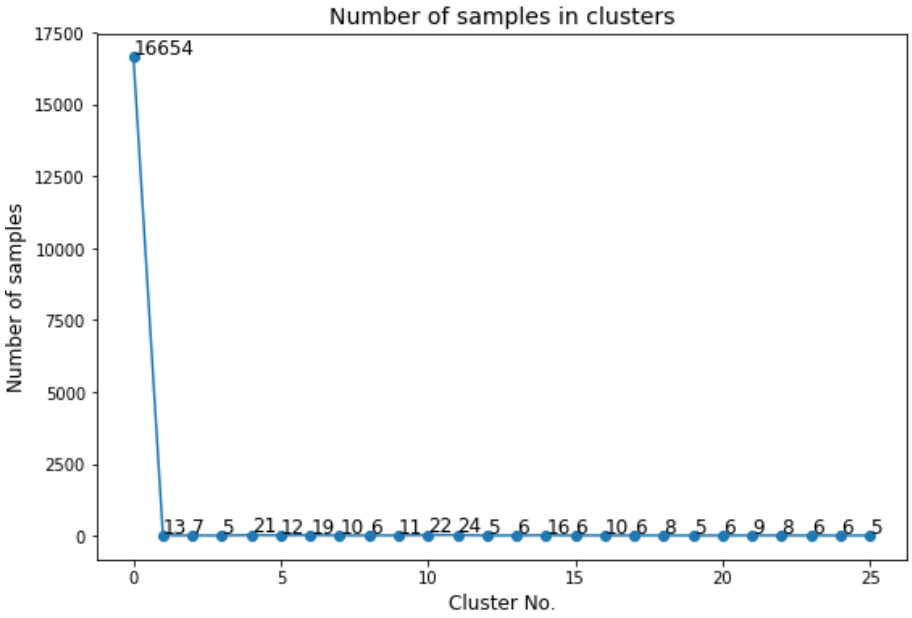

Figure 8-Size of clusters from DBSCAN for ECOINVENT with all bioflows

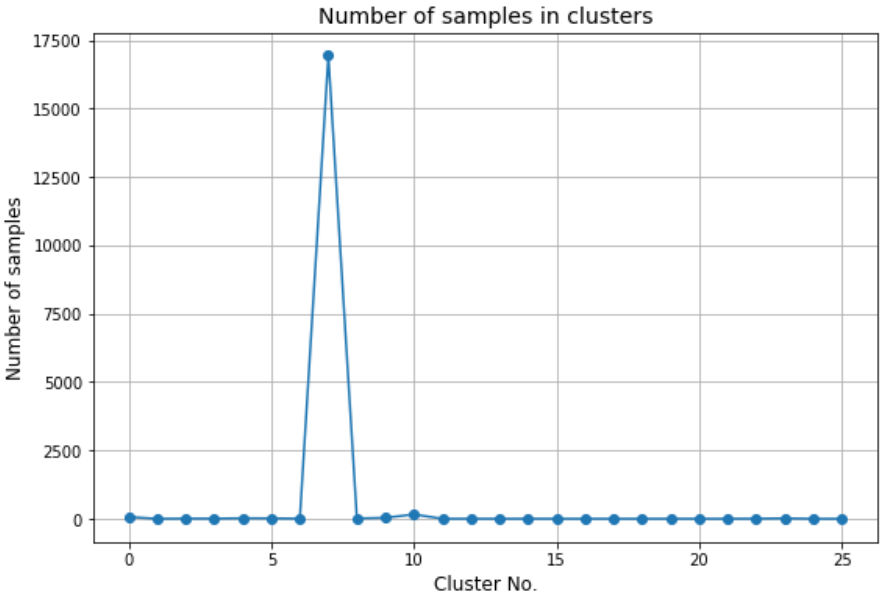

Figure 9-Size of clusters from Agglomerative Clustering for ECOINVENT with all bioflows

When applying elbow point strategy, 3 principle components can explain about 50% of the variance. Although 3 components are not enough from the perspective of cumulative explained variance, it's good for visualization. From Figure 10, it's clear to notice that most of the points are condensed in a core, while some other points are spread far from the core in a random way, which can be seen as outliers.

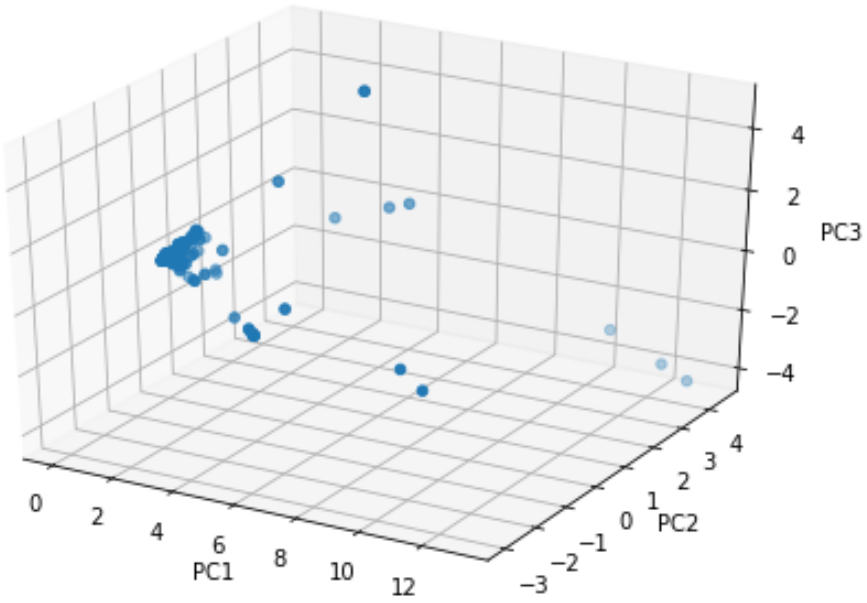

Figure 10-Plot of 3 principle components for ECOINVENT with all bioflows

For inventories with combined bioflows, which have similar name, after implement of PCA, the results are very similar, 61 principle components explain at least 90% of the information, elbow point is 4 and cumulative explained variance is around 55%, the results from different clustering algorithms are shown in Table 2. All the algorithms have very high Silhouette Coefficient scores, outliers are just an extremely small part, the other two indexes are not in the same levels, there is a huge difference regarding Calinski-Harabasz Index, and the Davies-Bouldin Index of DBSCAN is two times higher, but for the size distribution of clusters, they are in the same situation, exactly like the result for all bioflows as features, more than 90% of the samples are in one cluster.

Table 2-Clustering results for ECOINVENT with combined bioflows

Clustering algorithm	Number of clusters	Number of noise points	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
DBSCAN	7	84	0.978	1040.651	1.418
Agglomerative	23	0	0.976	138212.351	0.419
MeanShift	321	54	0.967	170.864	0.471

There is a way to reduce the number of bioflows further by retaining more important biolows and PCA. 17 ReCiPe Midpoint (E) V1.13 (ReCiPe) LCIA methods as shown in Table 3 are used to determine which bioflows are more important. For all the methods, if an elementary flow contributes less than 1 percent to the total LCA score for the activity, then it will be considered as an unimportant flow and be removed from the matrix. At the end, the number of features will be reduced to around 600. After removal of zero columns correlation analysis and PCA, the dataset is in a proper dimension for clustering.

Table 3-17 ReCiPe Midpoint LCIA Methods

Number	ReCiPe LCIA Method
1	('ReCiPe Midpoint (E) V1.13', 'freshwater ecotoxicity', 'FETPinf')
2	('ReCiPe Midpoint (E) V1.13', 'human toxicity', 'HTPinf')
3	('ReCiPe Midpoint (E) V1.13', 'marine ecotoxicity', 'METPinf')
4	('ReCiPe Midpoint (E) V1.13', 'terrestrial ecotoxicity', 'TETPinf')
5	('ReCiPe Midpoint (E) V1.13', 'metal depletion', 'MDP'),
6	('ReCiPe Midpoint (E) V1.13', 'agricultural land occupation', 'ALOP'),
7	('ReCiPe Midpoint (E) V1.13', 'climate change', 'GWP500')
8	('ReCiPe Midpoint (E) V1.13', 'fossil depletion', 'FDP')
9	('ReCiPe Midpoint (E) V1.13', 'freshwater eutrophication', 'FEP')
10	('ReCiPe Midpoint (E) V1.13', 'ionising radiation', 'IRP_HE')
11	('ReCiPe Midpoint (E) V1.13', 'marine eutrophication', 'MEP')
12	('ReCiPe Midpoint (E) V1.13', 'ozone depletion', 'ODPinf')
13	('ReCiPe Midpoint (E) V1.13', 'particulate matter formation', 'PMFP')
14	('ReCiPe Midpoint (E) V1.13', 'photochemical oxidant formation', 'POFP')
15	('ReCiPe Midpoint (E) V1.13', 'terrestrial acidification', 'TAP500')
16	('ReCiPe Midpoint (E) V1.13', 'urban land occupation', 'ULOP')
17	('ReCiPe Midpoint (E) V1.13', 'water depletion', 'WDP')

In order to retain at least 90% of the variance, 7 principle components should be used. And then these 7 principle components are used as features for clustering.

Table 4-Clustering results for ECOINVENT with important bioflows

Clustering algorithm	Number of clusters	Number of noise points	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
DBSCAN	16	48	0.192	80.629	1.211
Agglomerative	70	0	0.355	15039.258	0.861
MeanShift	13	2240	0.549	4990.574	0.805

From the results, DBSCAN gives the worst performance based on the three indexes. Although MeanShift gives higher Silhouette Coefficient and lower Davies-Bouldin Index, but there are over 2000 outliers. Meanwhile the sizes of clusters for DBSCAN and MeanShift are not as expected as shown in the Figure 11 and Figure 12 respectively. There is one cluster containing 80% or even more samples of the whole dataset as shown in the clustering results, which makes other clusters less important and thus the whole database can be seen as one cluster with some outliers.

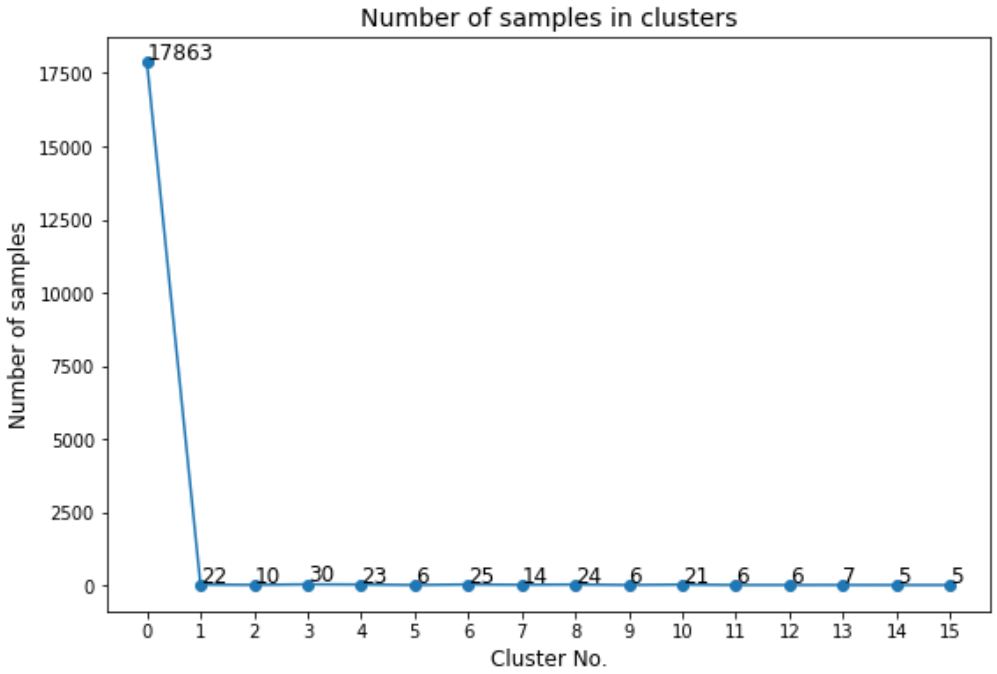

Figure 11-Size of clusters from DBSCAN for ECOINVENT with important bioflows

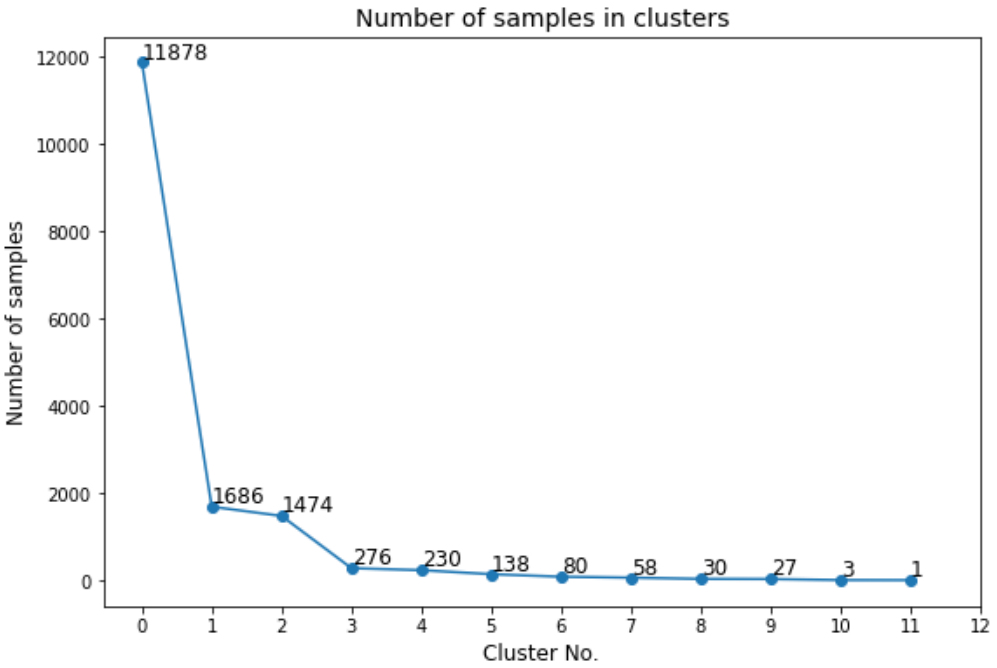

Figure 12-Size of clusters from MeanShift for ECOINVENT with important bioflows

Agglomerative Clustering gives a reasonable number of clusters and the size of clusters looks logical, as there are many clusters and the distribution regarding size of groups are much better than the results from DBSCAN and MeanShift as shown in Figure 13. However, when compared with ISIC and CPC, the activities belonging to the same ISIC or CPC group are distributed almost in every cluster like

randomness. There is no clue what are the interlinks between the activities in the same cluster, for example, electricity production from hard coal and electricity production from photovoltaic should not be in the same cluster, because the input features for clustering are the inventories from different activities, and the later has much less environmental impact than the former.

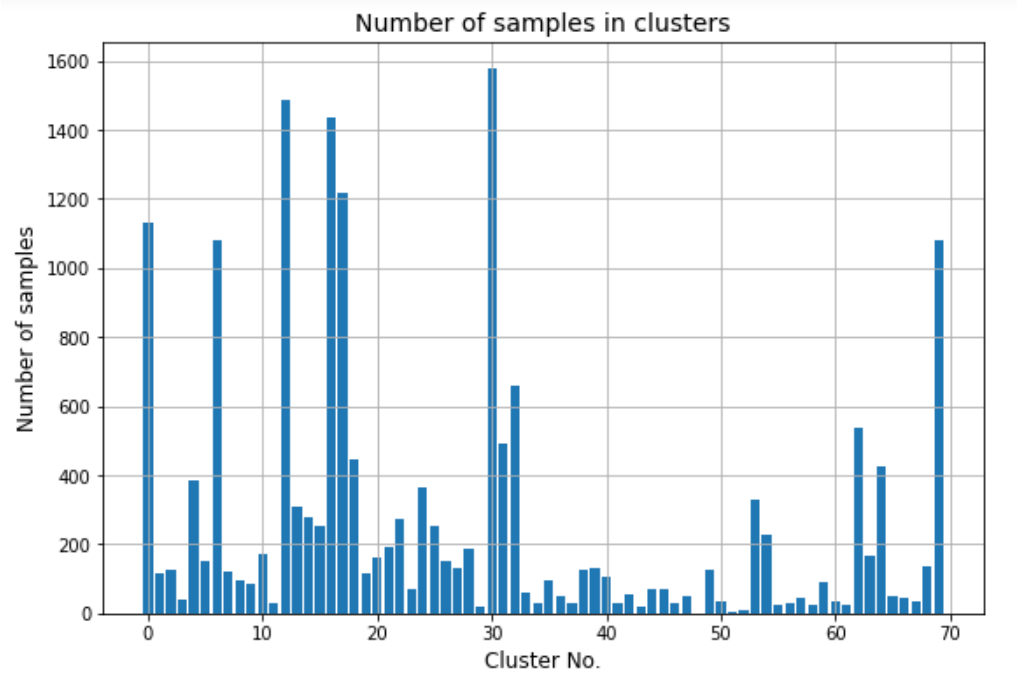

Figure 13-Clusters from AgglomerativeClustering for ECOINVENT with important bioflows

When using elbow point to select the number of principle components, the clustering results are similar. It is also possible to use features other than inventories for clustering, and LCA scores from ReCiPe LCIA methods would be an alternative. In order to test new features more efficiently it is better to focus on 'Market for' activities first, for each activity there will be an overall score under each ReCiPe LCIA method, so the reduced matrix will include 5397 samples and 17 features. It is important to apply feature scaling to make sure clustering measures are not affected by the different scales from different features.

The following Figure 14 shows the normalized LCA score for each market activity with different LCIA methods, most of the activities are close to each other as straight lines, except some have fluctuations, but basically they can be seen as one cluster. One reason for this might be the potential impacts of

different activities are in the same range for most of the LCIA methods, only a small part of the activities show a difference, and it's not enough for them to form an important cluster.

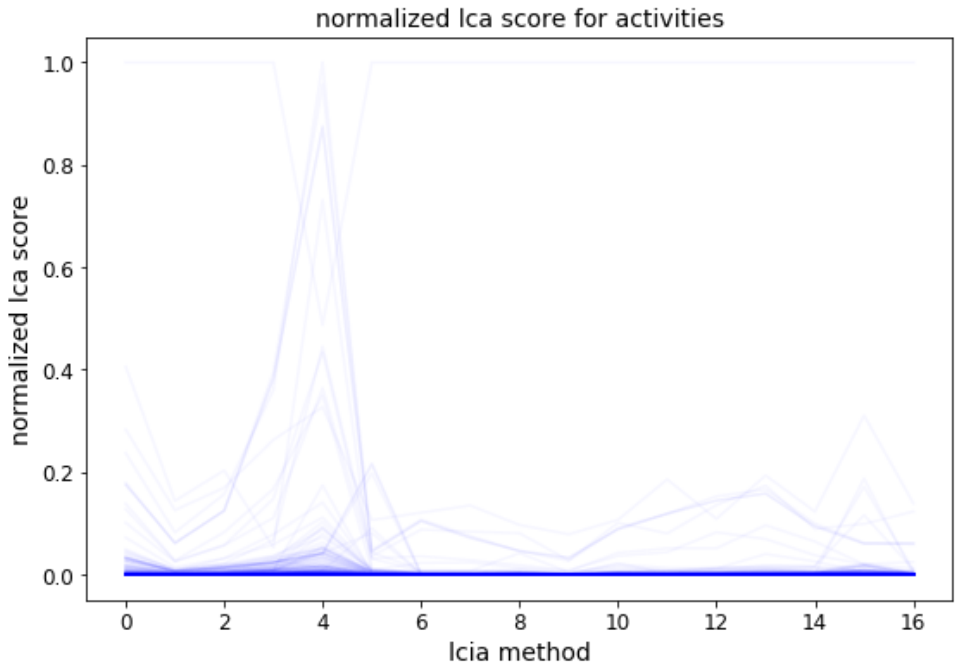

Figure 14-Normalized ReCiPe LCA scores for ECOINVENT market activities

And this can also be confirmed from the DBSCAN clustering algorithm, the result shows there is only 1 cluster with 12 outliers, and the Silhouette Coefficient is 0.996, which means the result is reliable.

There are 70 clusters while applying MeanShift method, but actually only 1 main cluster has much more activities (5297 out of 5397), and the other clusters normally include several samples (less than 5), which means the whole bunch of samples can be stored in one cluster, and there might be a high correlation between the LCIA methods.

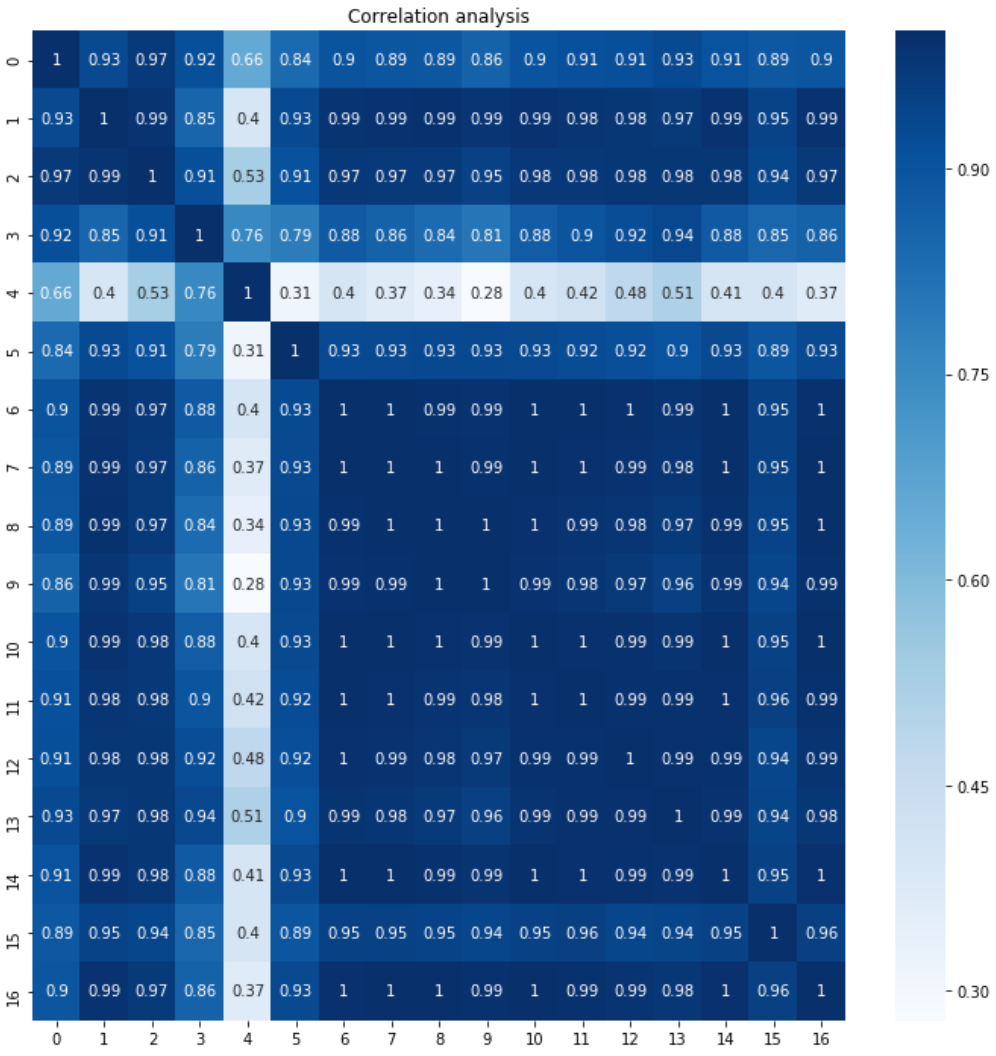

Figure 15-Correlation analysis for ReCiPe LCIA methods

The correlation analysis for the LCIA methods can be seen from Figure 15, as shown above, most of them have a high correlation between each other and the numbers are greater than 0.9 and even many of them are 1.0, except the ReCiPe ALOP method, which stands for 'agricultural land occupation'.

And this strategy also was applied to all the activities, the result is quite similar. As shown in the following Figure 16, most of the activities have the same LCA score as straight lines, while a small proportion of them have differences. When using the DBSCAN method, there is only 1 cluster with a few outliers (36 out of 18121), and the Silhouette Coefficient is 0.996. While applying Mean Shift, there would be 98 clusters, but only 1 cluster has a high number of activities (more than 17380), and the rest of them only have a few samples (less than 10). However, OPTICS didn't give a reasonable result as almost all the activities (more than 17000) detected as outliers.

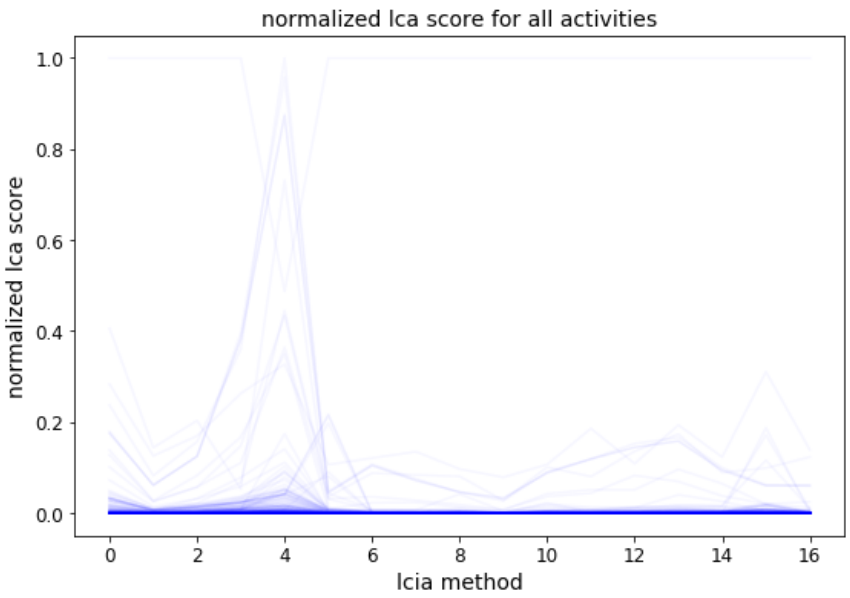

Figure 16-Normalized ReCiPe LCA scores for all ECOINVENT activities

These mean the scores from ReCiPe LCIA methods are not good features for activity clustering, because many LCIA methods are highly correlated between each other. Some other better features need to be found.

2.5.2 EXIOBASE

EXIOBASE has a much smaller structure with 7872 activities/processes and 58 bioflows. It is also easier to perform clustering algorithms. The following picture shows the real inventory values for each activity or process.

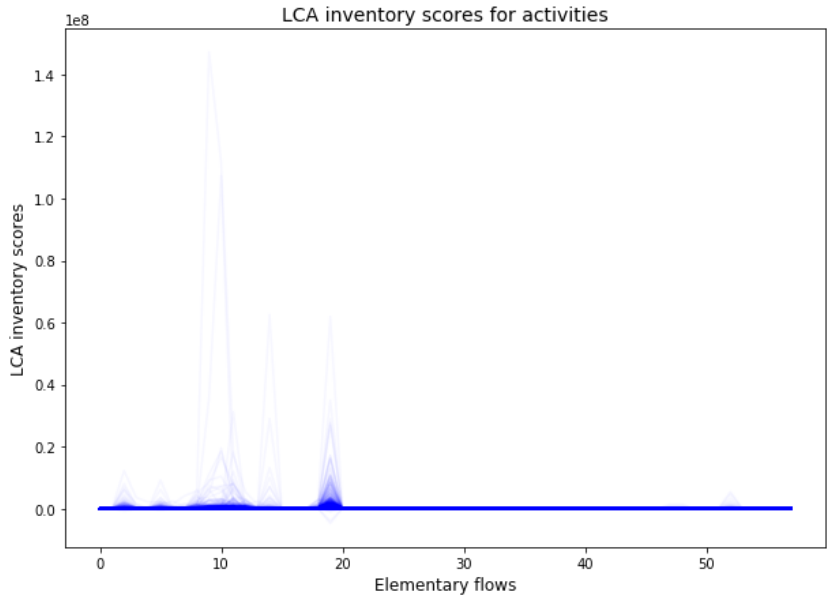

Figure 17-Initial inventories for activities in EXIOBASE

After feature scaling and correlation analysis, the number of features reduced to 48. As shown in Figure 15, there is a much clearer view of the features but still difficult to identify the hidden patterns.

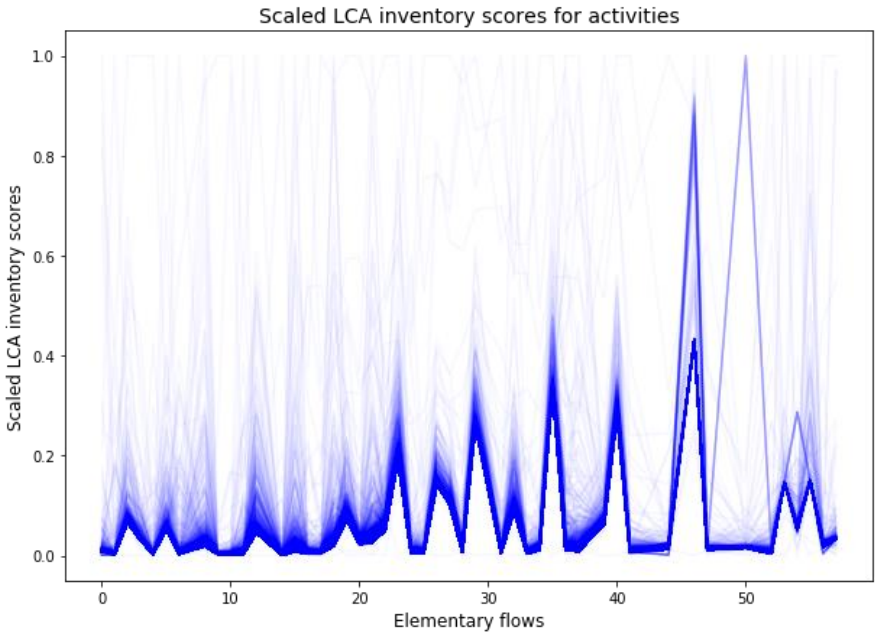

Figure 18-Scaled inventories for activities in EXIOBASE

When setting the scaled and uncorrelated inventories as features for clustering, the results from different algorithms are shown in the Table 5. Obviously OPTICS is not a good option as there are too many outliers and the performance indexes are the worst, while DBSCAN and Agglomerative Clustering look more reasonable, however, the size distribution of clusters are not as expected. Only 4 clusters from DBCSAN, but one cluster has over 7500 samples while the rest three have a total around 40. Although 35 clusters from Agglomerative, the largest one have more than 6700 samples, while most of the other clusters only have one sample.

Table 5-Clustering results for EXIOBASE with all bioflows

Clustering algorithm	Number of clusters	Number of noise points	Silhouette Coefficient	Calinski-Harabasz Index	Davies-Bouldin Index
DBSCAN	4	312	0.865	504.987	1.249
Agglomerative	35	0	0.726	2531.875	0.793
OPTICS	39	5190	-0.223	11.509	2.132

Since the number of elementary flows is limited, it doesn't make sense to combine or remove the less important flows to reduce the dimensionality. However, PCA was still applied for dimensionality reduction and visualization, no matter how much variance retained or how many principle components selected, no unexpected results appeared. Figure 19 shows a 3D plot of EXIOBASE activities after PCA with 3 PCs. Most of the points are clustered as a core, while others are spread like in the same plane but away from the core, there is also another small part of the activities formed a group above the main core, but too tiny to cause attention regarding size.

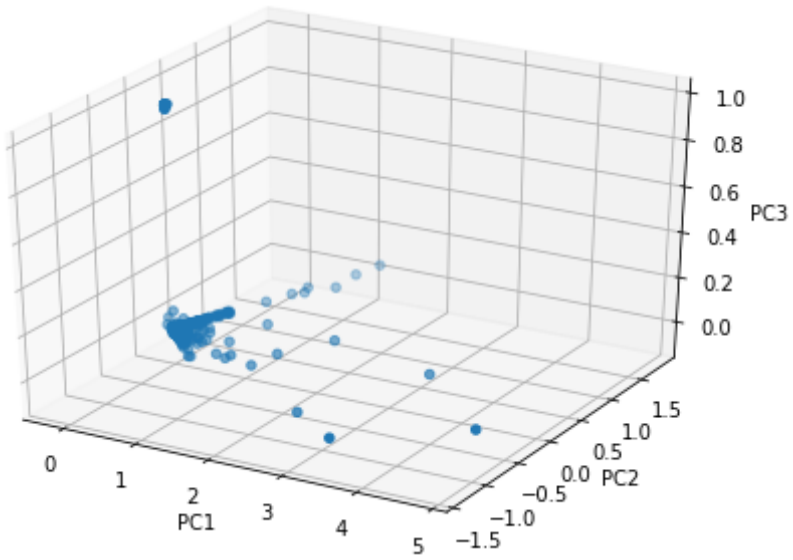

Figure 19-Plot of 3 principle components for EXIOBASE with all bioflows

As known from other studies and previous verification, some LCIA methods are highly correlated, so there is no surprise that when making LCA scores of 20 randomly picked LCIA methods as features, the result is similar.

It is always worthwhile to add new features when the initial ones are not working, locations of activities can be one. As shown in Figure 20, after adding the information of locations, the original features tend to vary in different directions at the end.

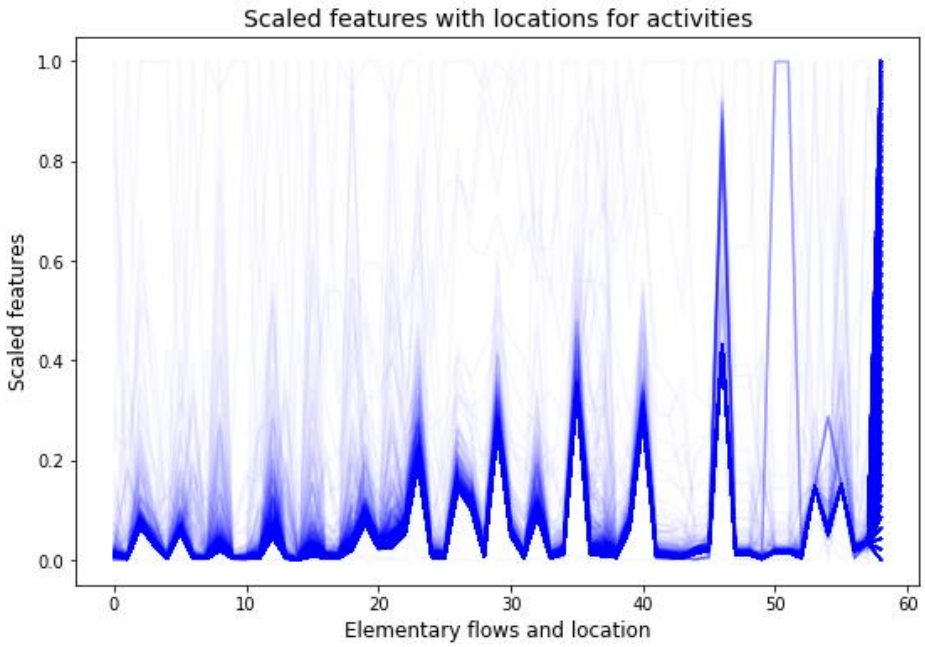

Figure 20-Scaled feature with locations for EXIOBASE

Although the clustering algorithms still do not give any desired outcomes, it is meaningful to notice that after adding new information, the dataset structure changed as shown in the Figure 21. Compared with Figure 19, the distributions of points are in totally different ways, even though from the perspective of density, the new dataset can still be considered as one cluster with some random outliers, the shape of the core samples changed significantly, which means proper new-added features can help with clustering.

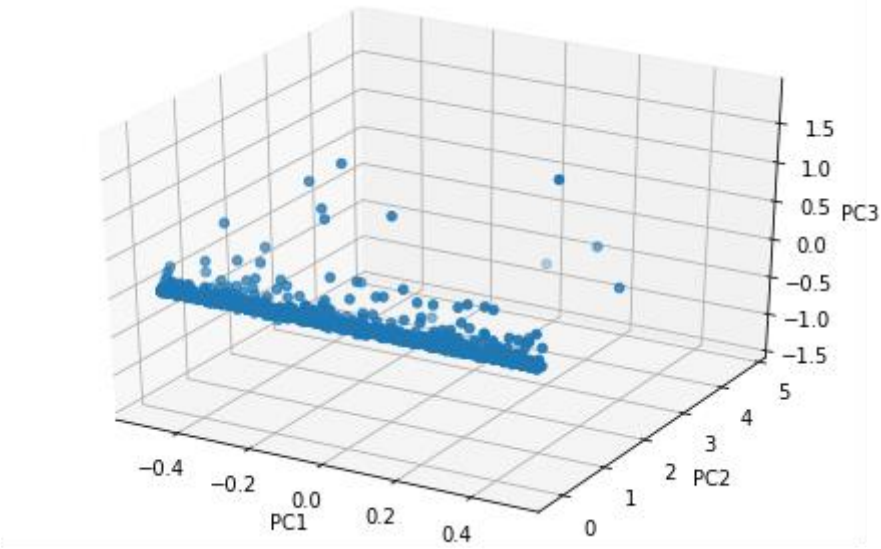

Figure 21-Plot of 3 principle components for EXIOBASE with locations

2.6 Discussion

With different dimensionality reduction techniques, input features as well as clustering algorithms being applied for ECOINVENT and EXIOBASE database, no hidden patterns or other useful information have been found yet. The possible reasons might be most of the activities in ECOINVENT or EXIOBASE, no matter it's about materials production or providing services, they all related with energy supply, in other words, energy is the basic requirement for the activities, therefore, when comes to the exchanges with bioflows, they may have similar values. Meanwhile most exchanges are pretty small, even though there are also large numbers, they don't have enough influence to change the overall patterns.

It's worth mentioning that dimensionality reduction from an engineering point might also be useful apart from mathematical methods. Adding proper new features will definitely improve the clustering results but it is not easy to find appropriate ones. Many LCIA methods are highly correlated, so maybe it's a good way to implement LCA calculations for certain LCIA methods and then use them as references for other categories of impact assessment and decision making.

Chapter 3

LCI estimation for wind turbines

LCI database is a very important information resource, however sometimes the data regarding some specific projects are not available, so it's necessary to develop a model for LCI estimation, specifically in this work the LCI estimation is for wind turbines, which is trying to estimate the detailed LCI for wind turbines with limited information of a wind turbine like rated power and location (onshore or offshore) by using machine learning algorithms as well as some other mathematical methods for LCA studies, it is also possible to predict the service time of a wind turbine and its total electricity production from the available datasets for better comparisons between projects.

3.1 The methodology

Renewable energy systems (RES) are being increasingly developed all over the world to weaken the dependence on fossil fuels and reduce the environmental impact for a sustainable future. Wind power, as an important part in renewable energy, has a tendency to grow rapidly, and wind turbines with larger and larger capacities are being installed.

Although wind farms can be built on renewable energy sources, it doesn't mean that wind energy would not have any negative impacts on environment, compared with the conventional fossil fuel based power plants, most of the environmental impacts come from the phases of manufacture and installation, where energy and materials are needed, while the emissions from operation phase are very limited.

Life cycle assessment (LCA) is a useful tool for environment impact assessment and decision making in energy systems. Meanwhile, LCA is also a very complicated method as it requires a lot of data for all the processes, however, not all of them are available, and a lot of ways have been implemented to solve this problem, like interpolation, using averages, or roughly estimation from engineering experiences. There is a parameterized model to generate life cycle inventory (LCI) for wind turbines in Denmark [25], which can generate LCI of wind turbines based on limited information and some mathematical estimations. As known, machine learning is a tool for predicting missing data, and therefore might be applied in LCA, specifically generating LCI data for wind turbines in this work.

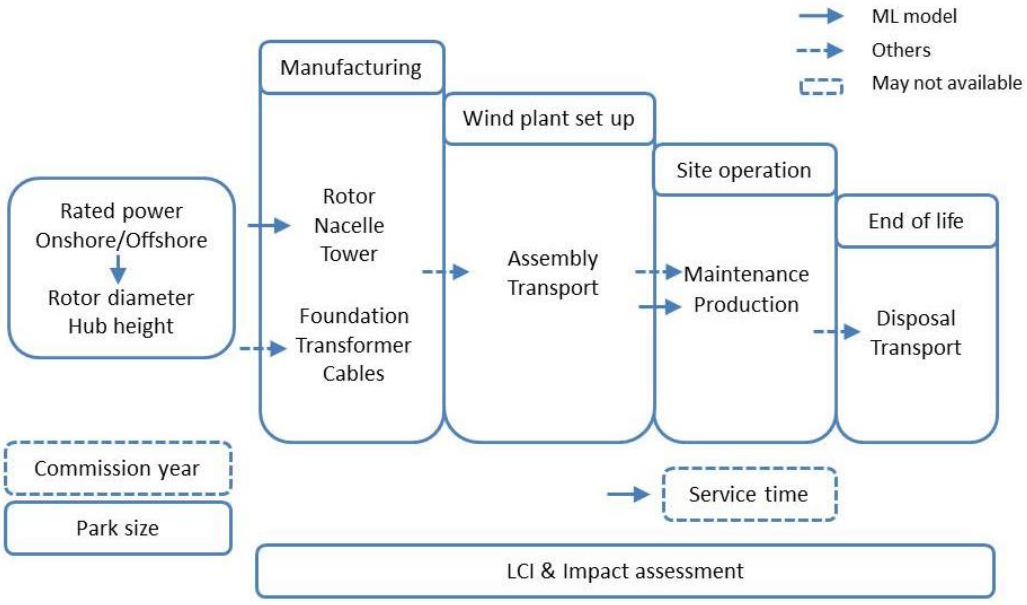

Figure 22-A graphic illustration of applying ML in estimating LCI of a wind turbine

A graphic representation of estimating LCI of a wind turbine by applying machine learning algorithms is shown in Figure 22. This model can be tailored based on the initial information of wind turbines, machine learning methods are implemented to predict the value of rotor diameter, hub height, masses of rotor, nacelle, tower, and then generate the LCI combined with other mathematical ways to size other components and activities of assembling, maintenance, transport and disposal by only knowing the nominal power and location. Generally, the model would be more accurate if there is more information available, and machine learning is there for missing values like hub height or mass of rotor. After estimating the LCI of a wind turbine, the next step is environmental impact assessment, which is based on an open source LCA software Brightway2 and ECOINVENT, a LCI database providing 'cradle-to-supply' inventories for different activities. Machine learning can also be applied in predicting service time and total electricity production, which gives information to calculate the impact in the same unit for better comparison with other projects.

3.2 Supervised machine learning

Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems [9]. When the desired output variable is continuous, then it is called a regression problem. There are many regression methods in machine learning can be applied to train the model and then estimate the some unknown values.

Linear regression typically involves a linear combination of input parameters, and it can be presented by

$$y(x, w) = w_0 + w_1 x_1 + \cdots + w_p x_p \quad (3)$$

Where x=(x1, x2, …, xp) is input variable, and w=(w1, w2, …, wp) is the vector of coefficients.

There are limitations of simple linear models, therefore we extend the models by polynomial processing to generate polynomials as inputs or by linear combinations of nonlinear functions of input variables.

Decision trees are a non-parametric supervised learning method, aims to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [26]. Decision trees usually learn from data to approximate a fitting curve with a set of if-then-else decision rules. Deeper tree has more complex decision rules and fits the model better. Random forest regression is an averaging algorithm based on decision trees.

Neural network, which is also known as Multi-layer Perceptron, is a supervised learning algorithm used for both classification and regression.

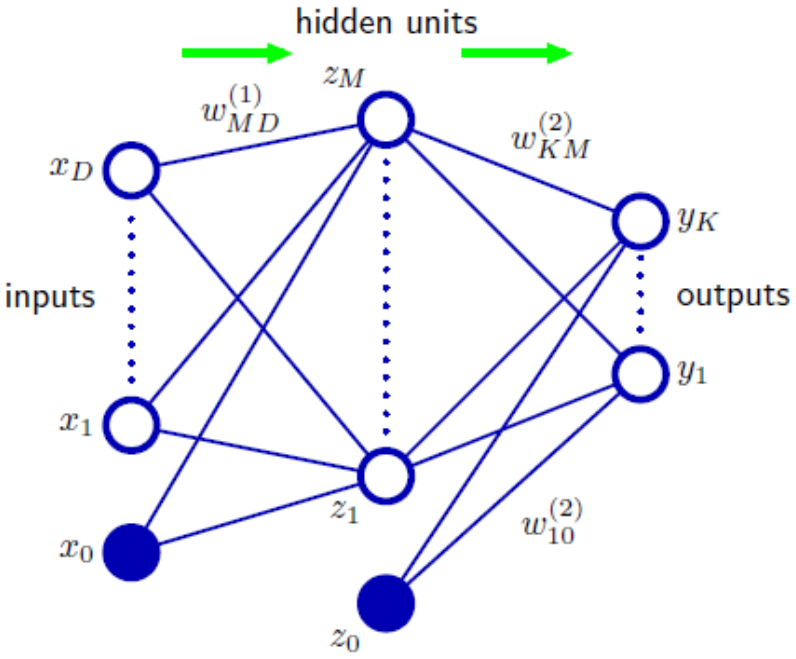

Figure 23-A representation of neural network diagram[9]

The input layer consists of a set of neurons known as input features. Every neuron in the hidden layer transforms the values from previous layer with a weighted linear summation, followed by an activation function, which is also called sigmoid function. The output layer gets the values and transforms them into outputs [26].

Support vector machine (SVM) [27] is considered as a powerful 'black box' learning algorithm in classification, regression and novelty detection, but more for classification problems, since it looks for a hyper-plane or set of hyper-planes to separate samples.

3.3 Model selection

Cost function is the residual sum of squares between the real targets in the dataset and the predicted values, which is used to find the best model by its minimum. However, it is not true to say the model with the lowest value of cost function is the best one, as there is a bias-variance dilemma, a few features might lead to under-fitting, which means a high bias, while too many features, especially polynomial ones would result in over-fitting, in this case it is high variance, and the value of cost

function may be extremely small, but it performs badly when predicting unknown targets. Therefore there should be a dataset for testing, usually takes 30% of the whole dataset. Meanwhile, a learning curve is also helpful to check how a model works regarding the bias-variance tradeoff. As shown in Figure 24, the error of testing of a good model usually decreases quickly with the increase of the training size, the error of training will go up a little with more training samples involved, but the difference between the errors of training and testing are getting smaller.

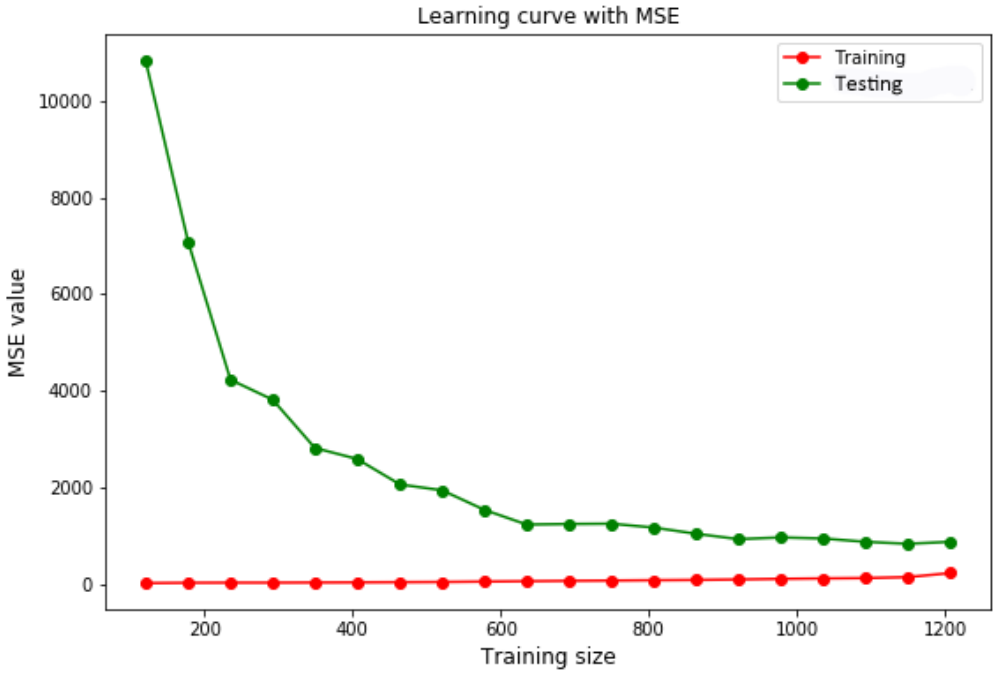

Figure 24-A learning curve for model selection

To compare different methods and evaluate their performance, four indexes are used for model selection.

-Training or testing score, which is the coefficient of determination R^2 of the prediction. The coefficient is defined as

$$R^2 = 1 - {}^{u}/_{v} \tag{4}$$

Where u is the residual sum of squares,

$$u = \sum(y_{true} - y_{pred})^2 \tag{5}$$

And v is the sum of squares of difference between true and mean values,

$$v = \sum(y_{true} - \overline{y_{true}})^2 \tag{6}$$

The best possible score is 1.0, which means the residual is 0.0 and normally this is not true case in practice, it can be negative when the model performs badly, and if a model always gives the mean of labels, the score would be 0.0.

- Mean absolute error

MAE is defined by,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=0}^{n-1}\left|y_{ture} - y_{pred}\right| \tag{7}$$

- Mean squared error and its root

MSE is defined by,

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=0}^{n-1}(y_{true} - y_{pred})^2 \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{9}$$

Based on their definitions, it's reliable to say that a better-performed model will have lower values of MAE, MSE and RMSE.

3.4 Machine learning models for LCI estimation

There is usually large amount of data involved in machine learning, so one challenge is gathering the data for training a model and testing. Fortunately, datasets about some dimensional parameters do exist, like rotor diameter, hub height, rated power of wind turbines, therefore, machine learning could be helpful when just using partial information to generate the life cycle inventory of a wind turbine.

In order to train a machine learning model for prediction, several steps are usually involved. Data collection is the first, and then it comes to data preparation, which is also known as data preprocessing, it requires data visualization to get insight in data, outlier detection, dimensionality reduction in case of high-dimensional problems, and feature scaling to ensure the data can be easily interpreted, next step is modeling with different machine learning algorithms and model evaluation by different performance indexes, if the trained models were not good enough, feature extraction and selection would be applied to create more features, modeling is more likely a iterative step, which requires model evaluation and selection. Finally, it's model interpretation, which aims to apply learned patterns or information for predictions.

From the engineering point of view, wind turbines with larger capacities have bigger dimensional size, which means the rotor diameter, hub height and masses of all the components are getting more when the rated power of a wind turbine is increasing, and there are some mathematical models for sizing

wind turbines [25]. Machine learning algorithms are applied to see if there will be a better result, and the model will be more general or not. For some parameters like rotor diameter, linear regression might be good enough to train the model, but other methods are also applied to make the comparison.

3.4.1 Rotor diameter

Suppose only limited information is available, if rated power and the location (onshore or offshore) of a wind turbine were known, is it possible to get the rotor diameter and how accurate it will be?

First of all, only with these two parameters, it's not precise enough. The following Figure 25 is a comparison between the predicted and actual values from a simple linear regression model. The training score is 0.81 and the testing score is 0.76. It's not good as expected, therefore, more features are needed for training the model, and these features can be derived from polynomial processing or feature extraction, which is to create new features based on the original ones.

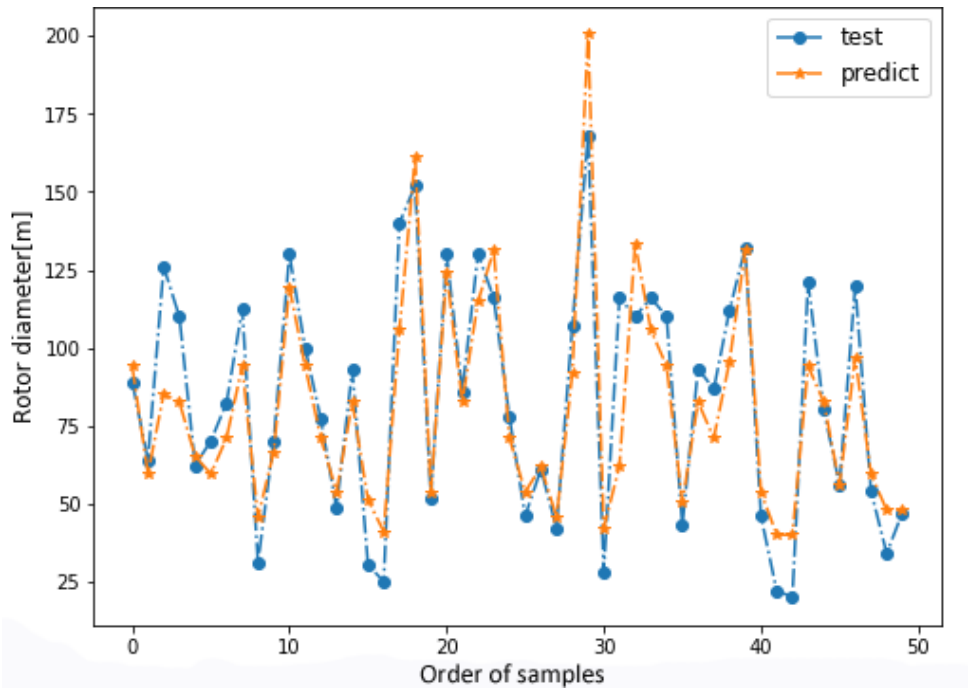

Figure 25-Comparison between predicted and actual values of rotor diameter

It is not true that we can get more accurate results with more features, there is a tradeoff between high bias and high variance.

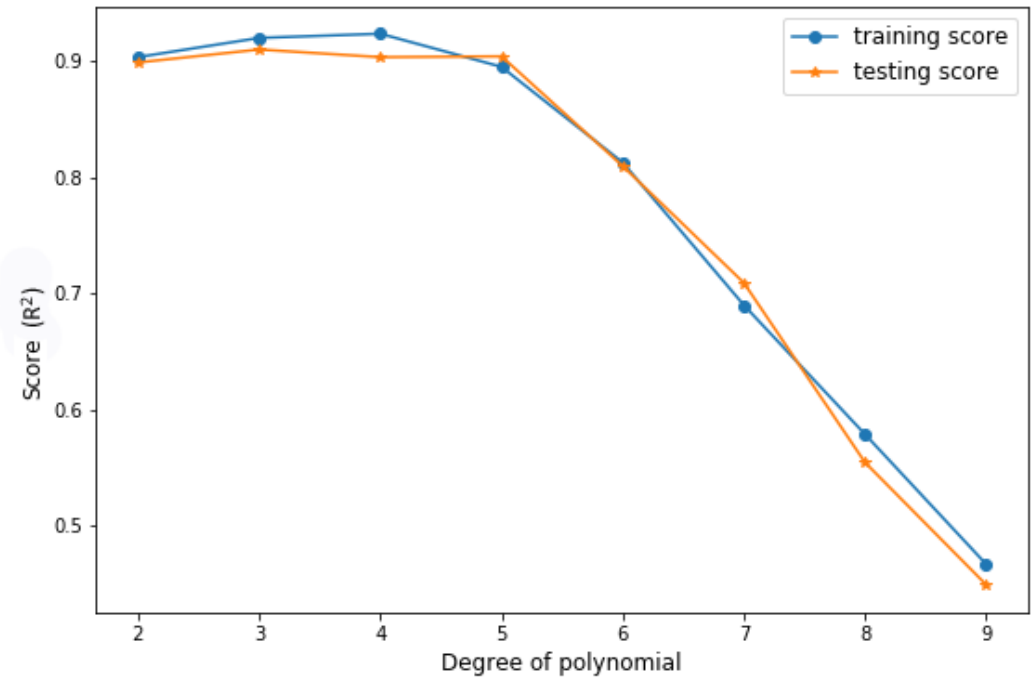

Figure 26-The performance of linear regression for rotor diameter with different polynomial degree

If we only have a few features, it is likely in a high bias situation (also named under-fitting), both the training score and testing score will be bad. And if we have a lot of parameters, it's possibly a high variance problem, in this case the training results are good, but when it comes to testing, the results vary a lot, and this is also called over fitting.

The Figure 26 above shows the performance of linear regression with different degree of polynomial features, with the increase of the polynomial degree, the training score and testing score also go up, however, when the degree is greater than 4, both training and testing score will decrease rapidly, which means the performance of the model gets worse with more features. In order to obtain a better result, only a proper number of features should be used.

As indicated in the Figure 26, when the degree is 3, the model has the highest testing score. Several machine learning regression models like linear regression, random forest, neural networks and support vector machines are tested with the polynomial features.

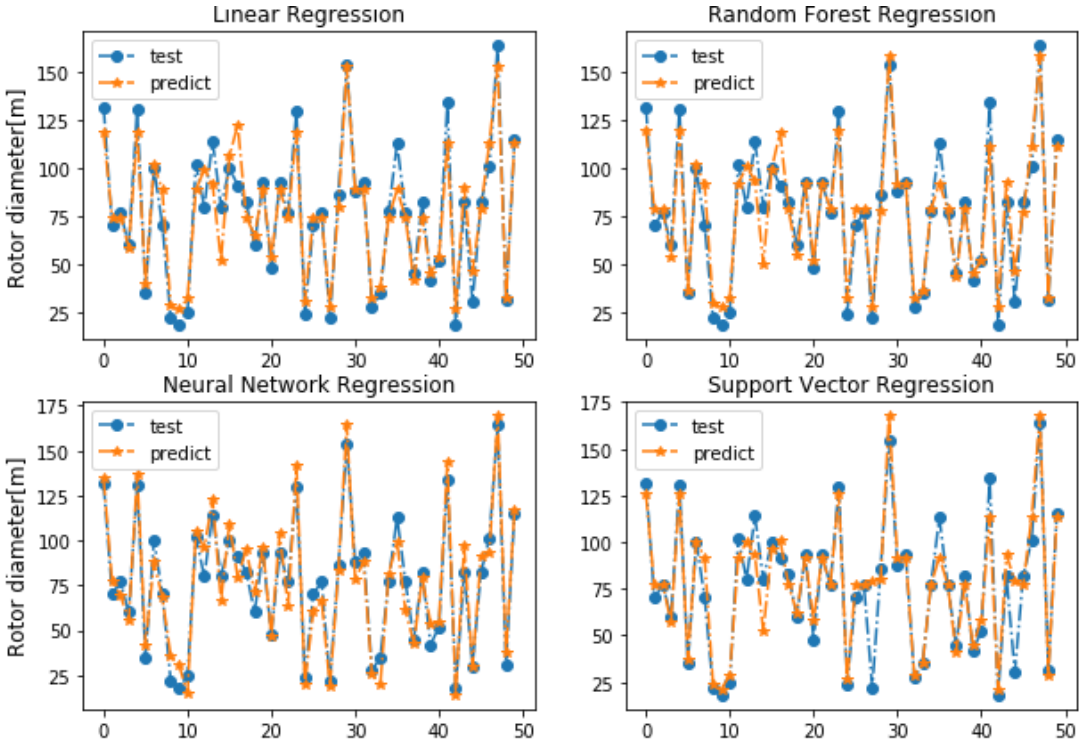

Figure 27-Different regression models for rotor diameter with polynomial features

As shown in the Figure 27, linear regression, random forest, support vector regression and neural network can train the model successfully and give good results. Linear regression is the easier way to perform, while random forest, neural network and support vector machine involve some parameters setting, and usually complex models may not give good results at all.

Table 6-Performance index of regression models for rotor diameter with polynomial features

	Training score	Testing score	MAE	MSE	RMSE
Linear regression	0.92	0.92	7.29	99.07	9.95
Random forest	0.91	0.92	7.68	105.06	10.25
Support vector	0.92	0.84	8.74	206.18	14.36
Neural network	0.87	0.82	9.02	141.85	11.91

Based on the results in Table 6, Linear regression has the highest testing score, MAE (Mean absolute error), MSE (Mean squared error) and RMSE (Root of mean squared error), it is better than the others, even though support vector regression has higher training score, but the testing score is much smaller.

Another way for creating more features is feature extraction, new features can be created from the existing features. In this case, we only know the location and rated power, so we can generate new features from rated power, like square of rated power, root of rated power and so on.

After creating new features based on rated power, then feature selection is applied to see which features are more important. There are many ways for feature selection, like removing features with low variance and recursive feature elimination. Different methods have their own principles and mechanisms, so there will be different results, usually it's better to perform at least two methods and try to find some features with common sense. It turns out that features including location, rated power, square root of rated power and natural logarithm of rated power are more crucial. The results from different methods are shown in Figure 28 and Table 7.

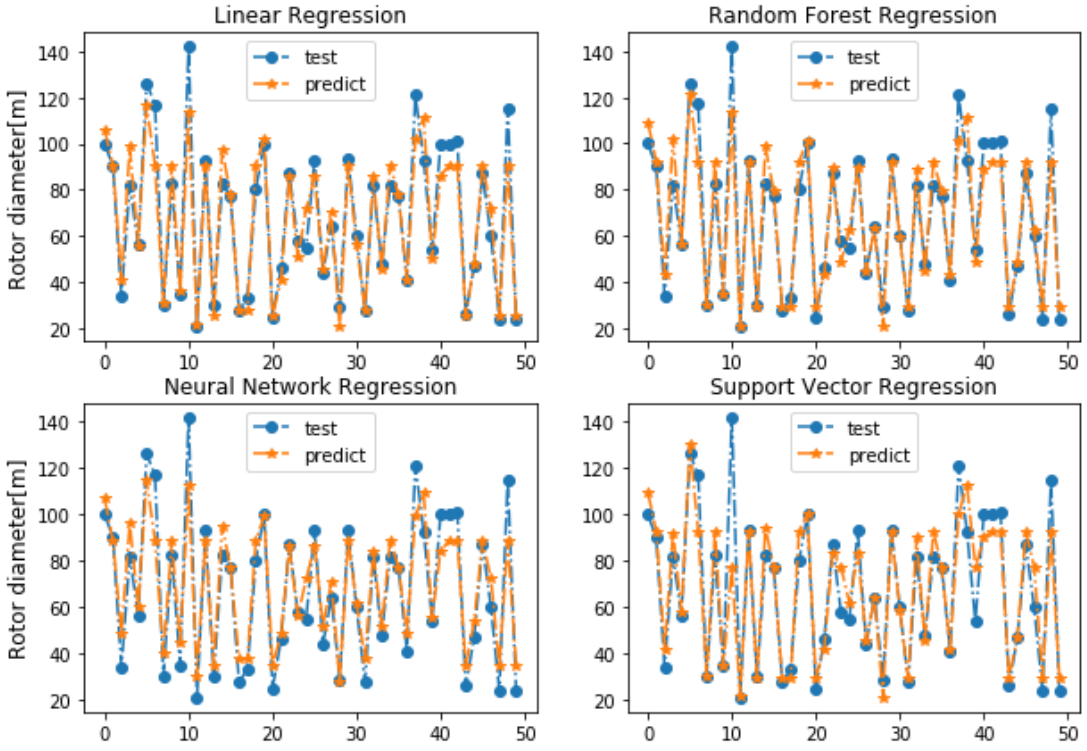

Figure 28-Different regression models for rotor diameter with selected features

Table 7-Performance index of regression models for rotor diameter with selected features

	Training score	Testing score	MAE	MSE	RMSE
Linear regression	0.91	0.93	7.09	94.60	9.73
Random forest	0.92	0.92	7.07	99.11	9.96
Support vector	0.92	0.88	7.74	154.81	12.44
Neural network	0.86	0.88	8.83	154.30	12.42

As shown in Table 7, linear regression still has the highest testing score, lowest MSE and RMSE, while random forest has the biggest training score and lowest MAE. Moreover, neural network performs much better than before this time, one possible reason for this is that fewer features are involved. To verify this, we only used the initial two features (location and rated power) to train the models and check their performances, however, the result is showing that neural network gives the worse result, and linear regression is just slightly better than neural network, random forest preforms better than others.

If there are a lot of features, it's better to do feature selection first to avoid over fitting, and neural network may not be a proper method for a problem with many features, linear regression and random forest are better. When there are just a few features and it's difficult to creating useful features, linear regression may no longer be appropriate, random forest and support vector regression are better, while neural network can also be applied.

3.4.2 Rotor weight

For rotor weight, we can use location, rated power and rotor diameter as original features. Just using these initial features, the prediction accuracy can be around 80%.

By using polynomial processing to create new features, the maximum degree is 2, as when it is bigger than 2, both the training and testing scores decrease as seen from Figure 29.

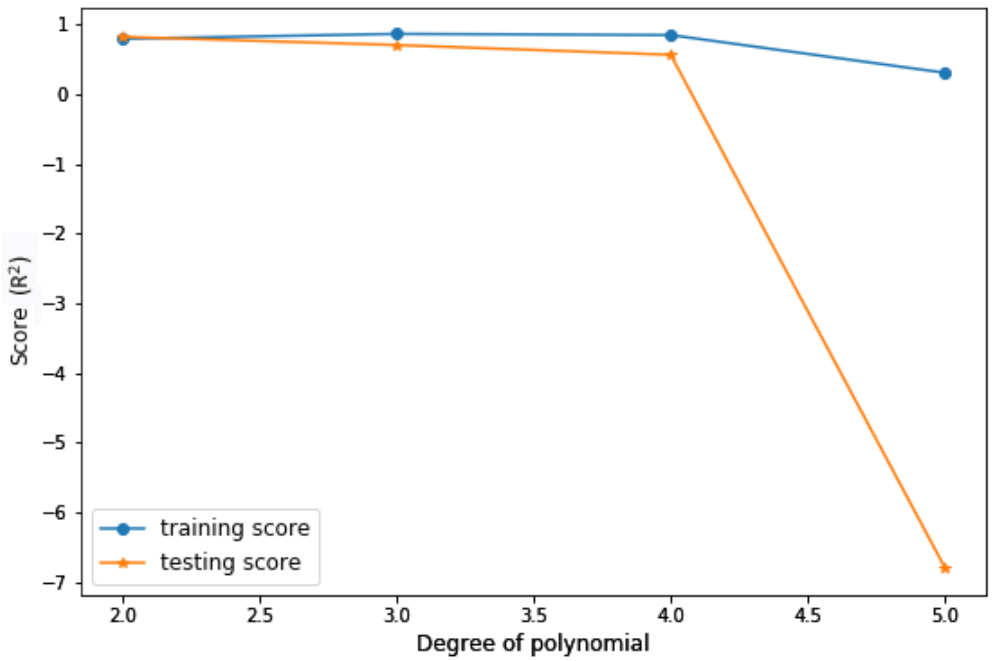

Figure 29-The performance of linear regression for rotor weight with different polynomial degree

Similarly, linear regression, random forest, neural network and support vector regression are applied for training the model for the features generated by polynomial processing with a degree of 2. And a comparison among these models is carried out according to Figure 30 and Table 8.

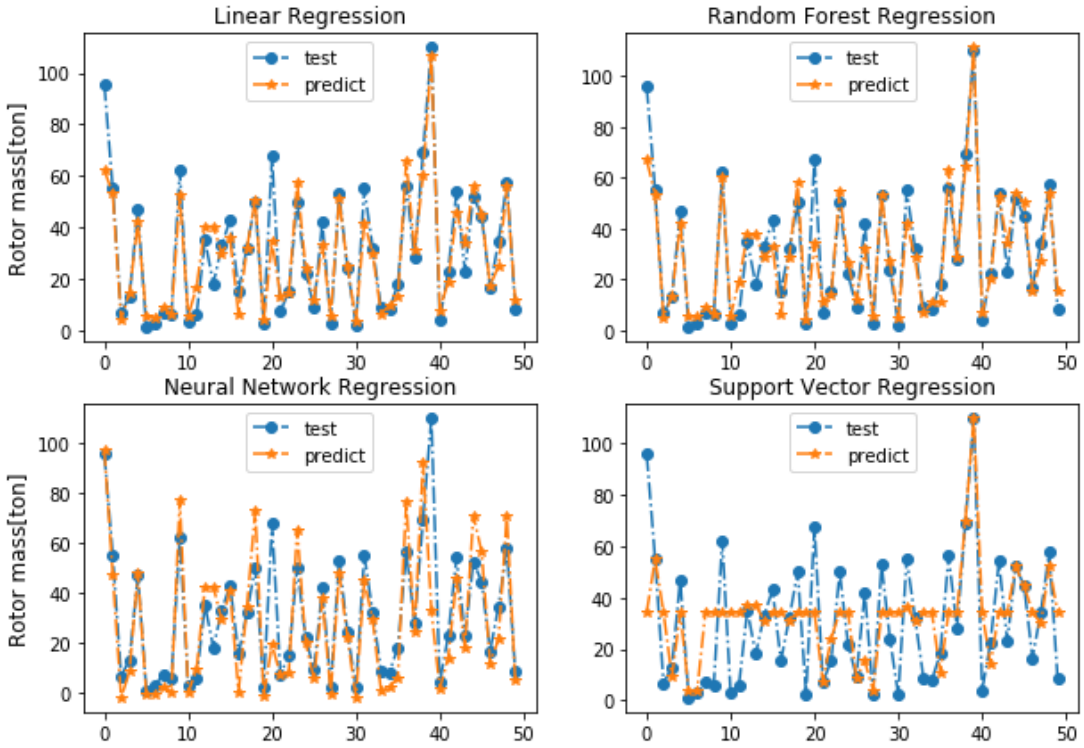

Figure 30-Different regression models for rotor weight with polynomial features

Table 8-Performance index of regression models for rotor weight with polynomial features

	Training score	Testing score	MAE	MSE	RMSE
Linear regression	0.82	0.77	5.83	174.80	13.22
Random forest	0.84	0.80	5.50	154.48	12.43
Support vector	0.90	0.26	13.78	573.86	23.96
Neural network	0.71	0.72	10.46	509.11	22.56

When using these features, random forest gives better results with the highest scores and lowest errors.

When applying feature extraction and feature selection, new features are created and used for training the models, the results are shown in Figure 31 and Table 9. Feature selection indicates that a combination of location, rated power and its square root, rotor diameter and its square will give better results. Random forest and linear regression are pretty close this time, they have similar validation

scores, linear regression has lower MSE and RMSE, while random forest has smaller MAE, however, the differences are tiny.

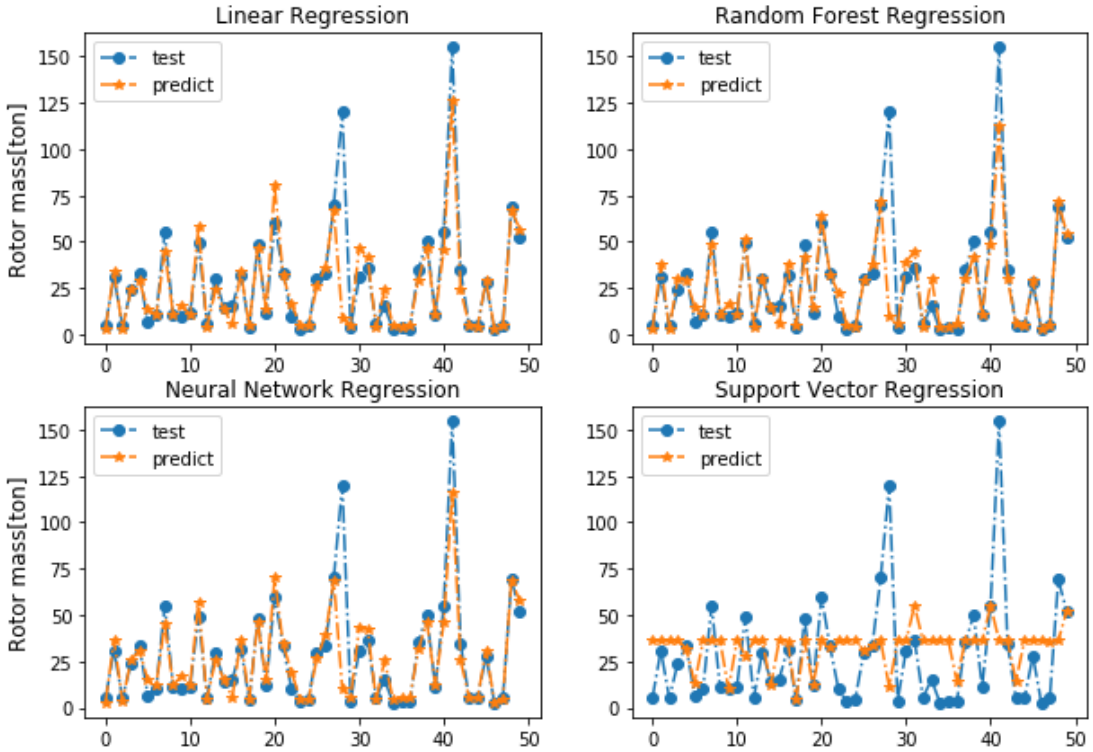

Figure 31-Different regression models for rotor weight with selected features

Table 9-Performance index of regression models for rotor weight with selected features

	Training score	Testing score	MAE	MSE	RMSE
Linear regression	0.79	0.87	5.19	66.03	8.13
Random forest	0.83	0.87	5.05	69.44	8.33
Support vector	0.90	0.29	13.60	389.84	19.74
Neural network	0.71	0.77	6.89	128.26	11.32

3.4.3 Nacelle weight

The initial available information is location and rated power. When only using this information, linear regression can give a result with a training score of 0.83 and a testing score of 0.81.

When applying polynomial processing, a degree of 2 gives the best scores, 4 different machine learning algorithms are applied with these polynomials as input features.

Table 10-Performance index of regression models for nacelle weight with polynomial features

	Training score	Testing score	MAE	MSE	RMSE
Linear regression	0.85	0.84	11.01	510.32	22.59
Random forest	0.82	0.90	10.40	311.72	17.66
Support vector	0.79	0.86	10.76	441.54	21.01
Neural network	0.48	0.74	24.24	828.16	28.78

As shown in Table 10 and Figure 32, random forest gives better results, higher testing scores and lower MAE and MSE values, while linear regression and support vector machine also give good results based on the overall performance.

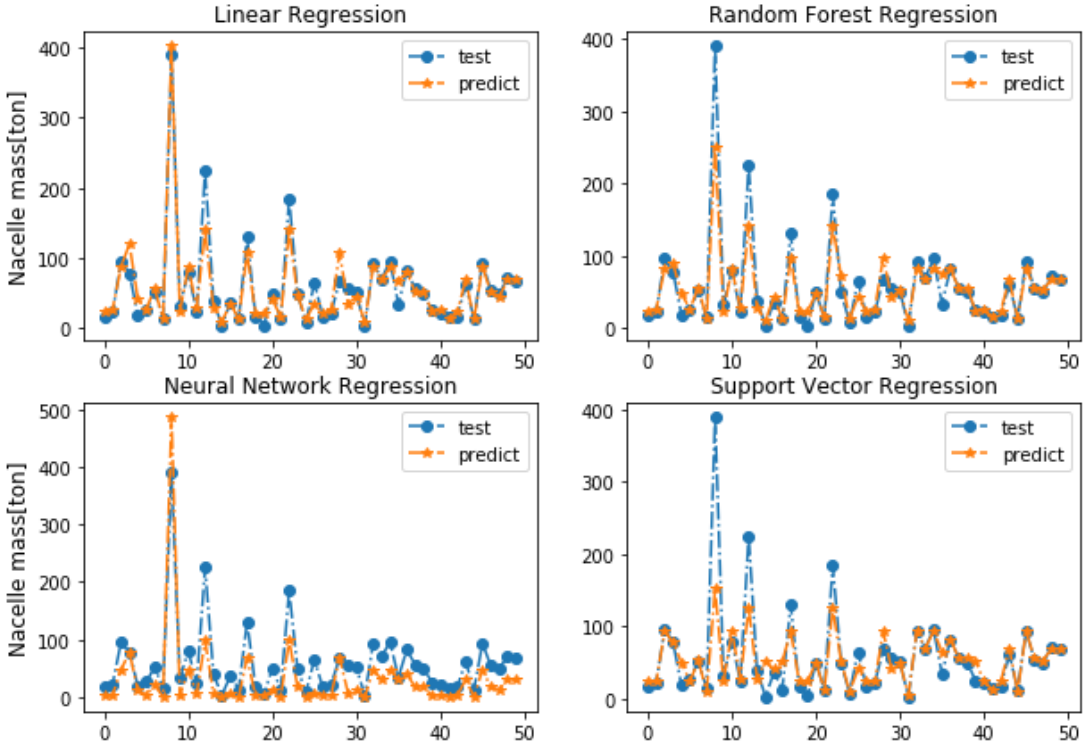

Figure 32-Different regression models for nacelle weight with polynomial features

With the features come from feature extraction and feature selection, input features are location, rated power, square root of rated power, and reciprocal of rated power. As shown in Figure 33 and Table 11, neither linear regression nor random forest gives the better result this time, the best model comes from support vector machine, its training score is 0.84, and testing score is 0.75, all errors like MAE, MSE

are the lowest, which demonstrates that linear regression or random forest regression is not possible to be better than others all the time.

Table 11-Performance index of regression models for nacelle weight with selected features

	Training score	Test score	MAE	MSE	RMSE
Linear regression	0.89	0.51	13.61	821.31	28.66
Random forest	0.88	0.62	11.39	628.79	25.08
Support vector	0.84	0.75	9.98	415.10	20.37
Neural network	0.69	0.23	27.38	1448.82	38.06

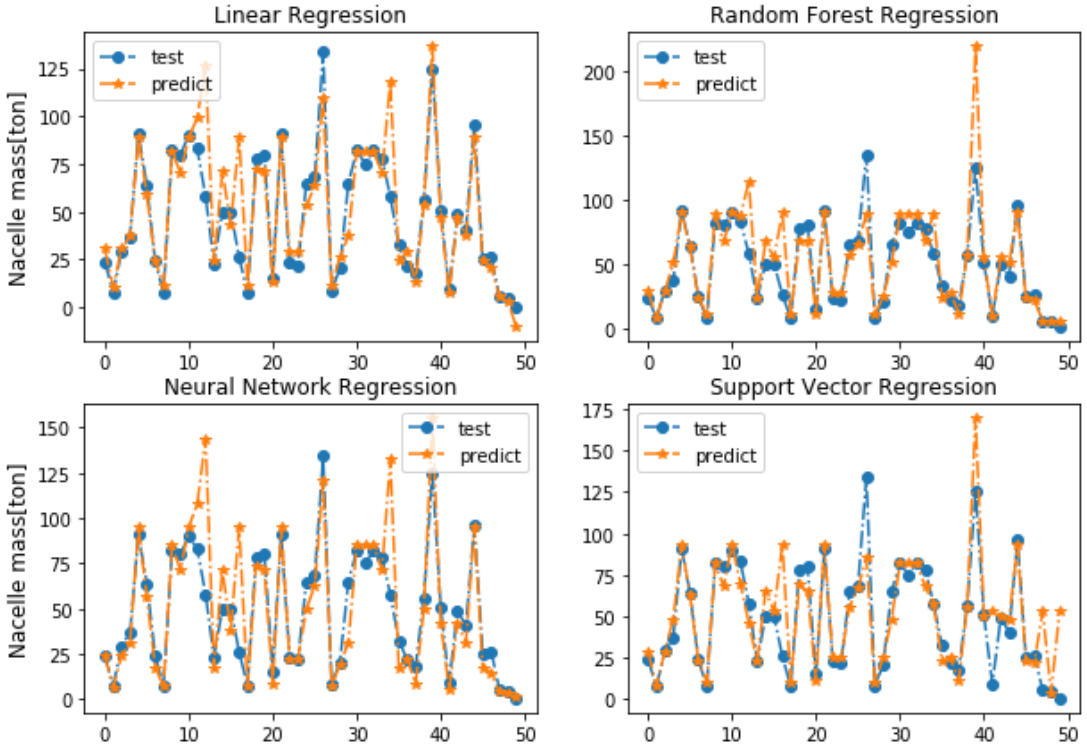

Figure 33-Different regression models for nacelle weight with selected features

3.4.4 Hub height

There are two datasets available for hub height of wind turbines. One dataset includes the location whether a wind turbine is offshore or not, while the other one doesn't. The common known features are rated power, rotor diameter.

When using the dataset with information about the location, so the initial features can be location, rated power, and rotor diameter. The highest testing score is around 0.82 from random forest.

With polynomial degree of 2, the best result comes from linear regression, the testing score is 0.82 and training score is 0.84. Support vector machine turns out to be over fitting, because the training score is extremely high, usually more than 0.90, but the testing score is below 0.60.

Regarding feature selection and feature extraction, the selected input features are location, rated power, rotor diameter, reciprocals of rated power and rotor diameter. The highest testing score is 0.82 from random forest.

Base on the results form feature selection, rated power and rotor diameter are better features than location. To verify this, another dataset without the information about the location was used. Dataset without information of locations consists of capacity, rotor diameter and hub height of wind turbines. Even if using a simple linear regression model with these two features, good result with both training and testing scores greater than 0.90 can be achieved.

When using features with a polynomial degree of 2, the scores can be greater than 0.95, and the best results come from random forest, which is 0.97.

With feature extraction and feature selection, the input features are rated power, rotor diameter, square root of rated power, natural logarithm and reciprocal of rotor diameter. The results from all the methods are greater than 0.90, and the best one is random forest, which is 0.97 for training and 0.97 for testing.

Table 12-Comparison among different dataset for hub height estimation

	Dataset 1 with location	Dataset 2 without location	Dataset 1&2
Num. of samples	1023	9122	10145
Max.	154.5	140	154.5
Min.	18	11	11
Mean	75.4	39.7	43.35
Median	77.0	32.5	39.25
Best model	RF	RF	RF
Training score	0.88	0.97	0.96
Testing score	0.84	0.97	0.96
MAE	8.11	1.91	2.78
RMSE	10.37	3.35	5.06

To make a more general and accurate model, it's better to combine these two datasets, and the result in Table 12 shows that the average value is bigger because there are large values from dataset 1, the MSE and RMSE increase slightly, while the training and testing scores only change a little.

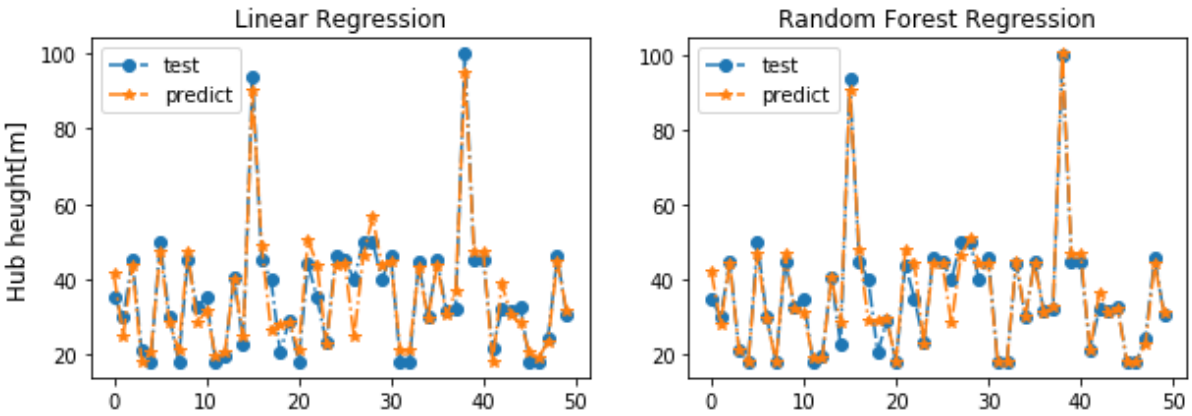

Figure 34-Different regression models for hub height with polynomial features

The models trained from the combined dataset are shown in Figure 34 and Table 13. As Neural Network and Support Vector Machine didn't give good results, the comparison is made without them. Both Linear Regression and Random Forest have good results, while random forest is better, although all the performance indexes become worse compared dataset 2, they are much better than dataset 1, which give better results for a more general and accurate model.

Table 13-Performance index of regression models for hub height with polynomial features

	Training score	Test score	MAE	MSE	RMSE
Linear regression	0.93	0.93	3.75	38.66	6.22
Random forest	0.96	0.96	2.67	24.59	4.96

3.4.5 Tower weight

There is a relatively smaller dataset for tower weight, the following graph shows the relationship between tower weight and other parameters, it seems there are some noise points as shown in Figure 35. And in practice, there are rarely wind turbine towers with a weight greater than 1000 tons.

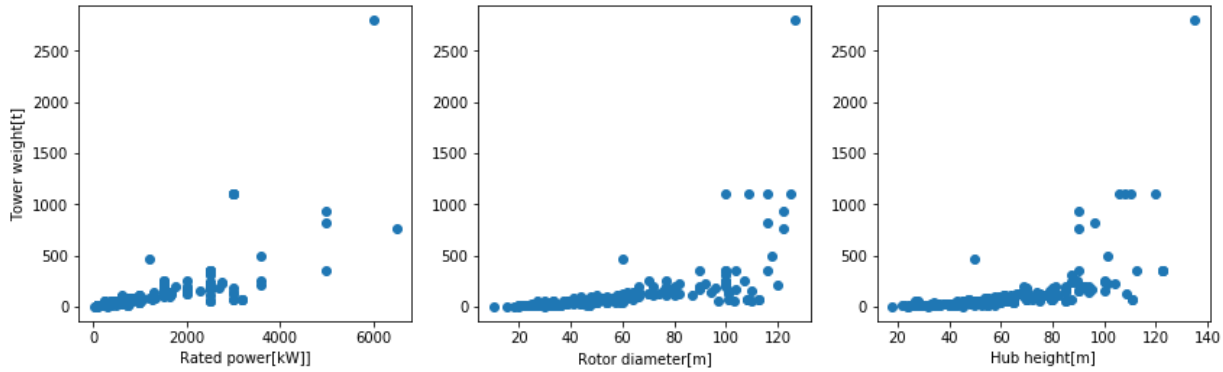

Figure 35-Visualization of tower weight with other parameters

When training the model without doing anything for the noise points, the result from different methods are shown in the Figure 36 and Table 14, they failed to predict large values over 1000 for both methods, the overall training or testing scores are not good enough or not good at all, since the MAE is around 60, which is not acceptable as the mean value is about 113, median is 56 for the whole dataset.

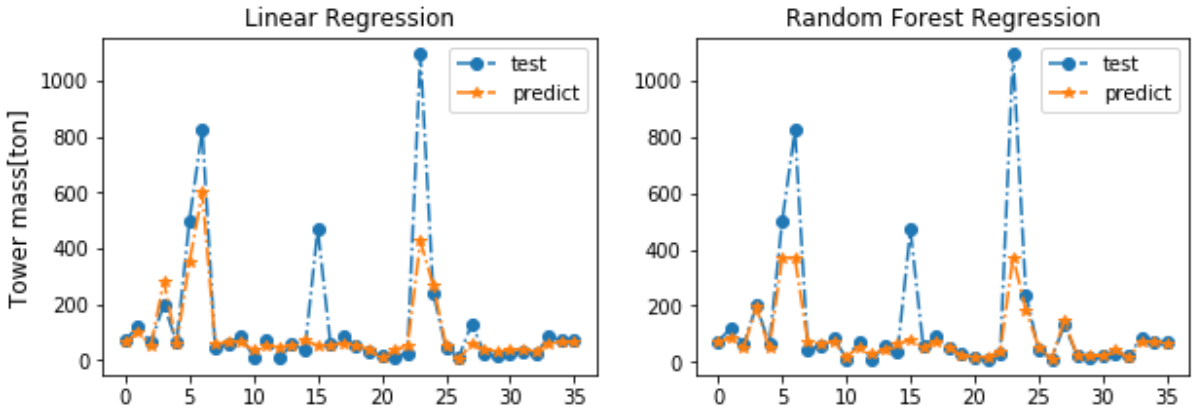

Figure 36-Different regression models for tower weight with outliers

Table 14- Performance index of regression models for tower weight with outliers

	Training score	Test score	MAE	MSE	RMSE
Linear regression	0.7061	0.8506	57.2638	19716.0517	140.4139
Random forest	0.5932	0.3431	61.7392	86710.3988	294.4663

One possible way to solve this issue is to adding more data, while another one is to remove the large values over 1000, which can be seen as outliers detection and it's easier to implement. Outliers are far away from the core of regular observations, which can be removed based on Gaussian distribution.

The following Figure 37 is a series of scatter plots between tower weight and other parameters like capacity, rotor diameter and hub height. Although there is no obvious trend and the data is in a wide range, but it seems reasonable as there is no strange points which are far away from the major part.

When using the new data to train the model, the performances improved a lot, as shown in Figure 38 and Table 15, first of all, no matter what kind of method used, both training and testing scores increased, MAE and RMSE reduced significantly compared before, the new mean and median values are 70, 53, the relative error reduced to 0.24 from 0.53 (MAE/mean).

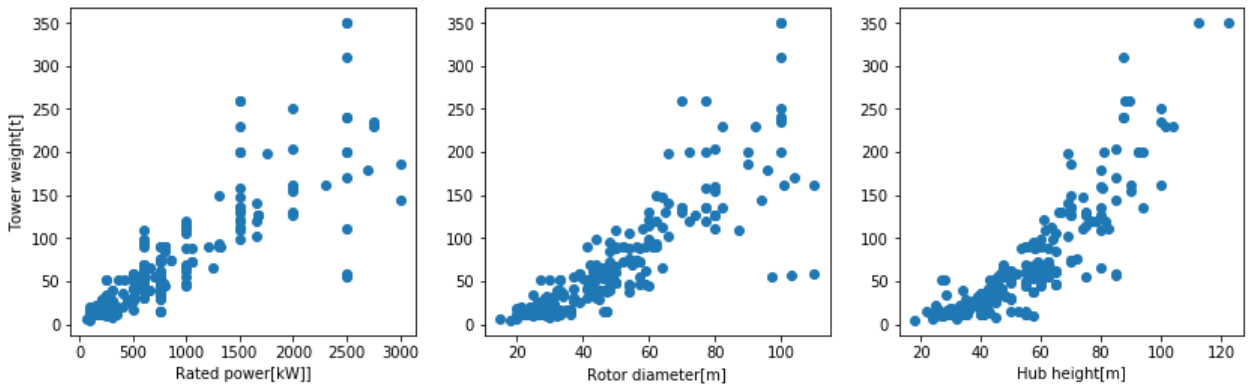

Figure 37-Visualization of tower weight with other parameters after removal of outliers

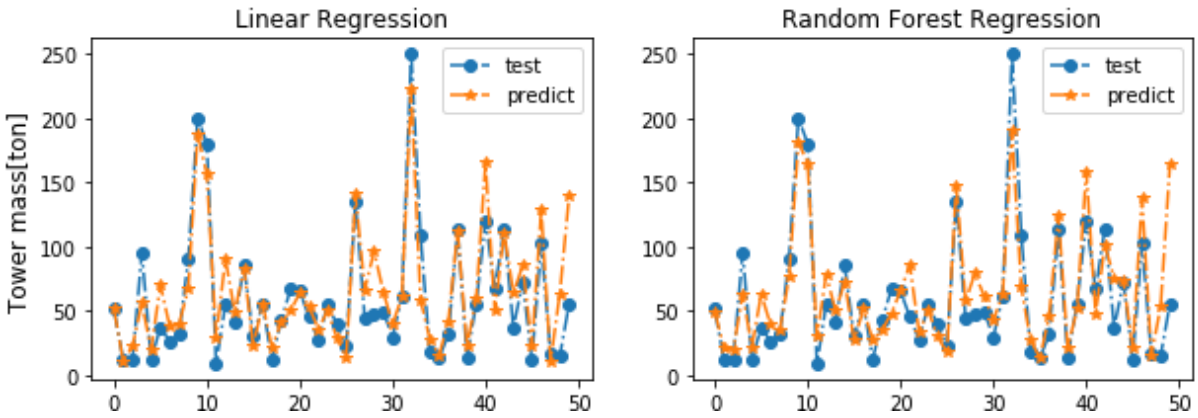

Figure 38-Different regression models for tower weight without outlier

Table 15-Performance index of regression models for tower weight without outliers

	Training score	Test score	MAE	MSE	RMSE
Linear regression	0.85	0.87	16.86	605.70	24.61
Random forest	0.84	0.82	17.29	833.37	28.87

Feature extraction and feature selection were also applied to create and select important features, but the trained models didn't give any better results, the highest validation score is about 0.77 for both linear regression and random forest. If more accuracy is required then more efforts needed for the feature creation.

3.4.6 Life time

From the histogram of Figure 39, we can see the life time of wind turbines from the dataset has a normal-looked distribution, with a mean value of 18.83, standard deviation of 4.83, most values are in the range of 10-30 years.

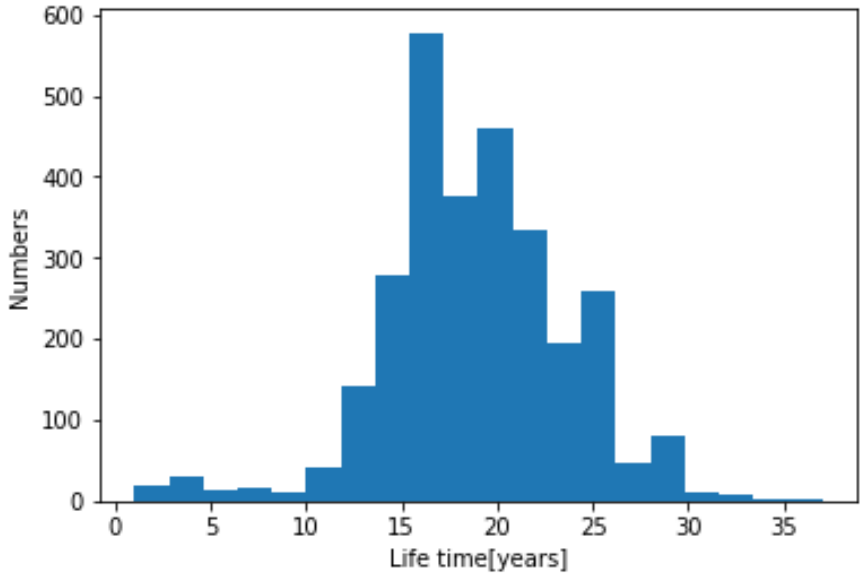

Figure 39-Histogram of life time for wind turbines

Typically, we use life time of wind turbines as labels, use rated power, rotor diameter, hub height, starting year as features. The best result is from Random Forest with a testing score of 0.57. As shown in the following Figure 40, most of the values are from 15 to 26, and there are also some numbers out of the main range varying from 4 to 14, if not considering them, the predicted data is like a normal distribution. However, the result is not satisfying.

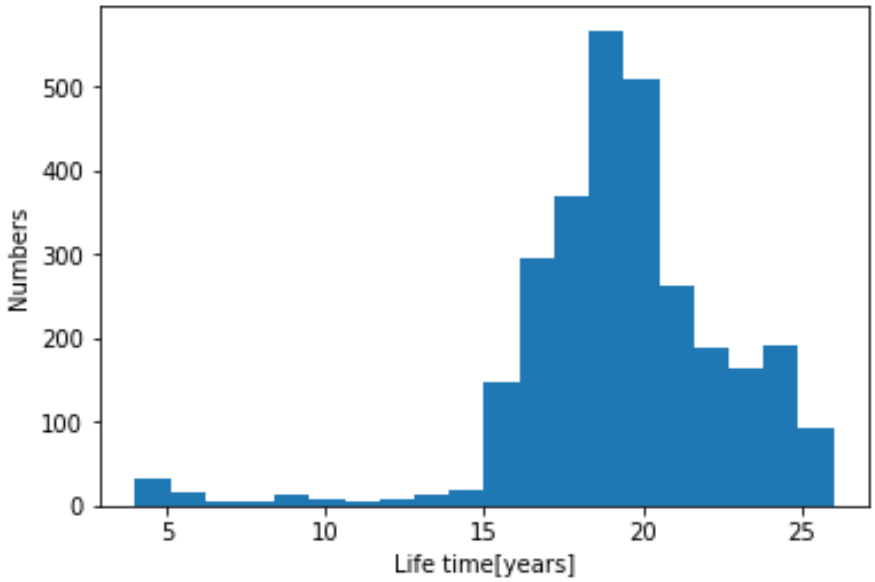

Figure 40-Histogram of predicted life time with life time as label

When setting the end year as labels, the service time range becomes wider as shown in Figure 41, the performance of the model improved a little bit, the testing score is higher, the values of MAE and RMSE are lower, but the differences are not huge. While the estimations from these two methods have similar mean and median values compared with the actual values, the distributions of life time estimated by these two methods have a similar shape, but totally different from the original dataset.

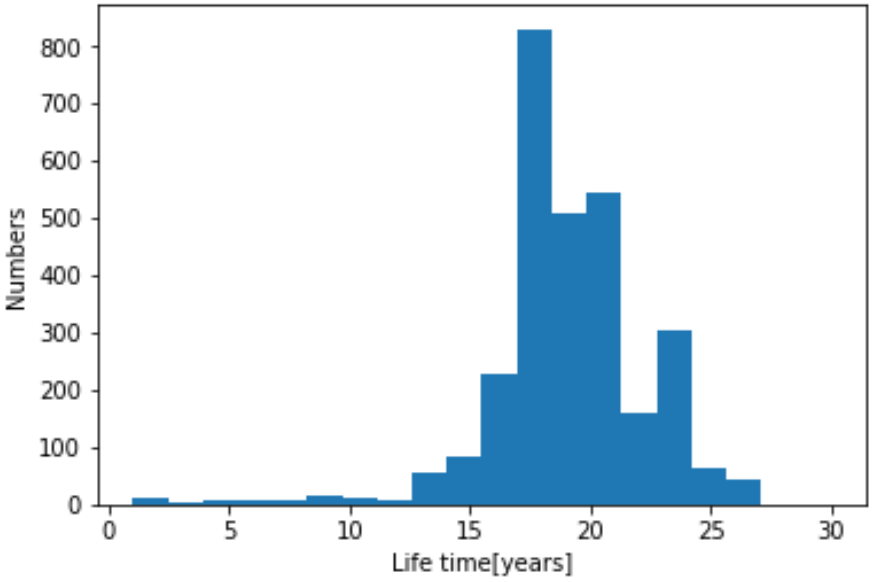

Figure 41-Histogram of predicted life time with end year as label

Although the results from both ways are not good enough, they are still better and more reasonable compared with the way that just setting 20 years as life time for all wind turbines, both MAE and RMSE are smaller, which means the predicted results are closer to the actual values as indicated in Table 16.

Table 16-Comparison among different scenarios for life time estimation

	Actual	Life time of 20 years	Life time as label	End year as label
Max	37	20	26	29
Min	1	20	4	2
Mean	18.83	20	18.87	18.51
Median	18	20	19	18
Best model	-	-	RF	RF
Training score	-	-	0.61	0.67
Testing score	-	-	0.57	0.64
MAE	-	3.95	2.32	2.10
RMSE	-	5.19	3.45	3.20
MAE/Mean	-	0.20	0.12	0.11

3.4.7 Electricity production

The purpose of estimating the total electricity production is to bring the environment impact assessment to the same level, usually on a basis of per kWh electricity produced. In most LCA reports of wind turbines, they estimate the total electricity production based on the service time of 20 years and a fixed annual capacity factor, which might be working but not accurate.

There is a dataset from Danish Ministry of Energy [28] consisting of enough information to train a model with machine learning and predict the electricity production of a wind turbine during its whole life time. The input features include rated power, rotor diameter, hub height, and the summation of electricity generated in each year from commissioning to decommissioning would be the target values.

When including life time as a feature, it's possible to obtain a model with higher testing score, lower MAE and RMSE values by polynomial processing with a degree of 2. As shown in Figure 42 and Table 17, Random Forest gives the better result, although the MAE is a big number, the testing score is still around 90%, this is because the total electricity production of a wind turbine is usually a much larger amount.

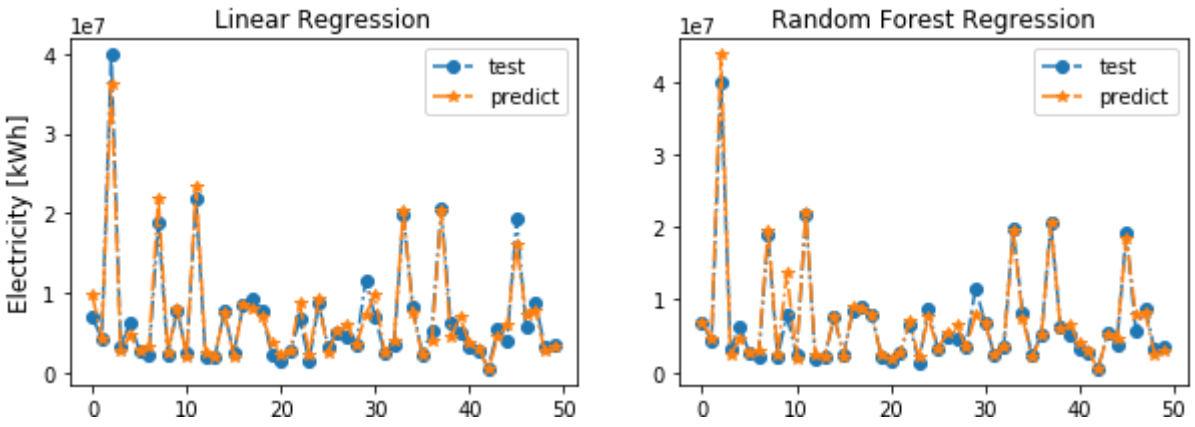

Figure 42-Total electricity production estimated by LR and RF with life time

Table 17-Performance index of regression models for electricity production with life time

	Training score	Test score	MAE	MSE	RMSE
Linear regression	0.72	0.75	2023323.0	1.19e13	3449916.5
Random forest	0.90	0.90	1041812.0	6.14e12	2478049.0

There are also some other ways to estimate the overall electricity production of a wind turbine, like using capacity factor or average annual wind speed.

When using the capacity factor, the total production of electricity can be defined by

$$\text{Electricity production} = \sum_{i=\text{start year}}^{\text{end year}} 8760 * P * Cf_i \quad (10)$$

Capacity factor is the ratio between the actual electricity energy output and the maximum possible electricity output during the same period. Average annual capacity factor varies year from year, the values of load factor for wind turbines in the past years can be found from https://www.renewables.ninja/, a website provides information for renewables like wind, solar energy.

Table 18-Average annual capacity factor [%] in Denmark from 1985 to 2016 [29]

Year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Onshore	23.29	27.01	23.98	27.29	27.55	29.31	26.35	26.88	28.15	29.23	27.47
Offshore	32.77	37.36	33.25	37.87	36.72	38.98	35.53	36.75	38.01	39.43	37.58
Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Onshore	24.34	25.50	28.39	25.01	27.34	23.64	25.24	22.46	25.81	25.49	23.37
Offshore	33.37	34.55	39.36	34.40	36.84	32.67	35.32	31.43	35.76	34.93	32.59
Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
Onshore	28.01	26.53	24.09	22.61	27.10	26.33	24.44	25.83	29.16	23.86	
Offshore	37.84	36.22	33.93	32.08	37.51	36.53	33.54	36.30	39.38	32.96	

However, the data for average annual wind speed is not available yet, otherwise there is a fast estimation [30] of electricity production would be applied for yearly production but the service time is also required. Then the comparison mainly made between the trained random forest regression model and equation defined above.

Table 19-Comparison of electricity production between RF prediction and CF calculation

	Actual	RF prediction	CF calculation
Max	101164434.2	44876274.9	74811276.0
Min	101.0	95584.1	20901.4
Mean	6934353.4	6975754.5	8522015.0
MAE		1041811.9	1901035.1
RMSE		2478049.0	3277604.0

As shown Table 19, actual values have a much bigger maximum and much lower minimum, while the minimum may be may be removed from the dataset since it's too small to be the total electricity production of a wind turbine. Random forest prediction gives a better result, the mean value is much closer, the MAE and RMSE errors are smaller, which indicates that the overall estimation made by RF regression model have lower differences, and the following Figure 43 gives a better understanding when comparing these two estimation methods. The values are picked randomly from the dataset, the overall values from capacity factor (CF) calculation are normally much bigger than actual numbers, while predicted ones are also larger, but the gaps between actual values and RF prediction are generally smaller than the gaps between actual numbers and CF calculation, therefore, the RF prediction is a better way to estimate the total electricity production than CF calculation.

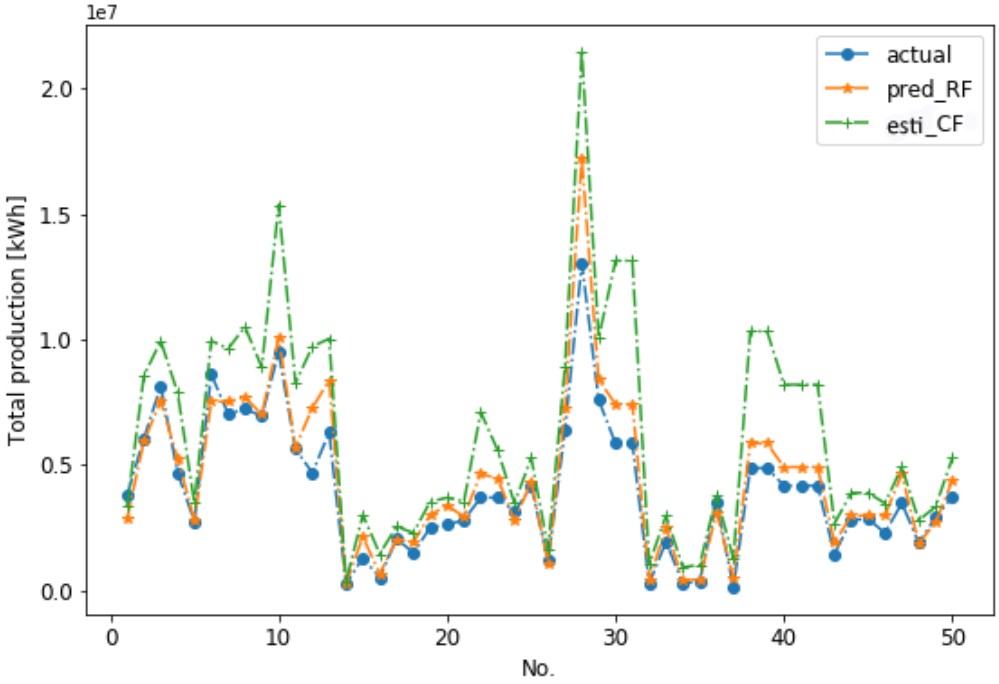

Figure 43-Comparison of electricity estimation with life time among different models

However, some information of wind turbines like construction year, service time may not be available, and the machine learning models for predicting the life time didn't perform well. In this case, it's also possible to train a machine learning model without using such data like life time. The following Figure 44 shows a result from Linear Regression and Random Forest after applying feature extraction and feature selection. Compared with the models using life time data, the overall performance is not as good as previous, both testing scores of Linear Regression and Random Forest decreased, the MAE and RMSE errors went up. And the better result is from Random Forest, its testing score is around 0.84 according to Table 20.

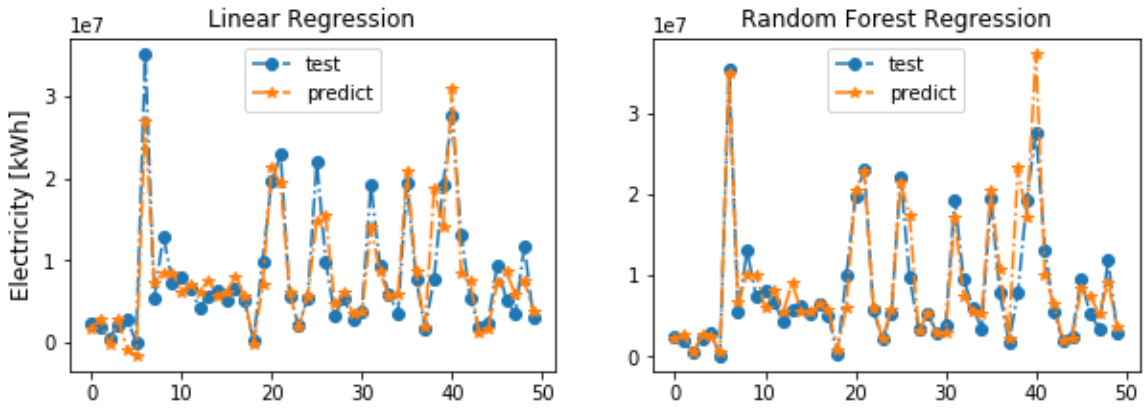

Figure 44-Total electricity production estimated by LR and RF without life time

Table 20-Performance index of regression models for electricity production without life time

	Training score	Test score	MAE	MSE	RMSE
Linear regression	0.79	0.74	2142085.1	1.41e13	3750164.7
Random forest	0.85	0.84	1380884.8	9.43e12	3071621.3

Although the performance of models trained without using life time shows they are inferior methods, but it's still a promising way when the input data is insufficient, as shown in the following Figure 45, the differences between RF prediction and actual values become larger without using life time as an input feature, however, the gaps are still smaller than the ones between CF calculations and actual numbers. The MAE and RMSE are still much smaller than CF calculations.

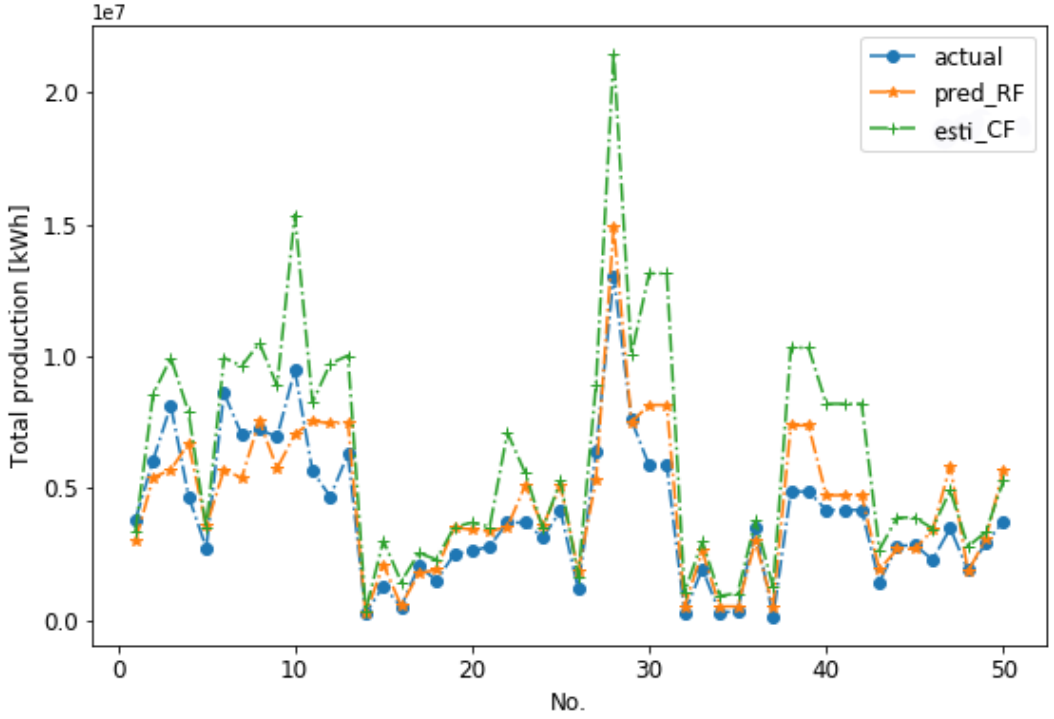

Figure 45-Comparison of electricity estimation without life time among different models

3.5 Other components and materials breakdown

With enough available data, some parts of a wind turbine can be sized by machine learning methods. While other parts like foundation, cables and transformers can't be estimated by machine learning as lack of data. Therefore, other models need to be applied.

3.5.1 Foundation

Foundations of onshore wind turbines mainly consist of concrete and reinforcing steel, the weight of foundations is calculated considering a sizing proportional to the tipping moment [31]. It is good for wind turbines with high capacity and high hub height, but the estimation gap will increase for low capacity and low hub height wind turbines as shown in Table 21. As wind turbines with higher capacity and hub height will be installed in the future, it is not necessary to modify the model just for wind turbines with low capacities.

Table 21-Weight of foundations calculated for different turbines

Capacity(kW)	Rotor Diameter(m)	Hub Height(m)	Foundation Weight(t)	Calculated Weight(t)	Error*
5000	126	80	2693	2692.5696	0.02%
5000	126	100	3367	3365.712	0.04%
5000	126	125	4208	4207.14	0.02%
5000	126	150	5050	5048.568	0.03%
3000	100	80	1696	1696	0.00%
3000	100	100	2121	2120	0.05%
3000	100	125	2651	2650	0.04%
3000	100	150	3181	3180	0.03%
800	50	50	238.4	265	-11.16%
600	43	40	191.2	156.7952	17.99%
150	24	30	53.4	36.6336	31.40%

*Error is calculated by (Foundation weight – Calculated weight)/Foundation weight

Monopile foundations are by far the most common type of foundation and have been used for 70-80 % of all offshore Wind Turbine Generators (WTGs) in operation today [32]. Therefore, sizing the foundation of offshore wind turbines is based on a Monopile one. The main components are pile, transition piece, grout and aggregates for scour protection as shown in Figure 46.

As shown in the following Table 22, they are average values for components of different offshore wind turbines [32], the capacity is from 3 MW to 10 MW. The material for pile and transition piece is steel, grout is made of cement with an average density of 1650 kg/m3, and aggregates are more like sands or rocks, which is not taking into consideration. Simple regression is used to obtain the values of each part for wind turbines with different capacities.

Table 22-Data for Monopile foundations of different wind turbines [31]

Capacity [kW]	3000	3600	4000	8000	10000
Pile weight [t]	400	475	550	725	900
Transition piece [t]	200	200	210	285	335
Grout [m3]	25	25	30	42.5	47.5

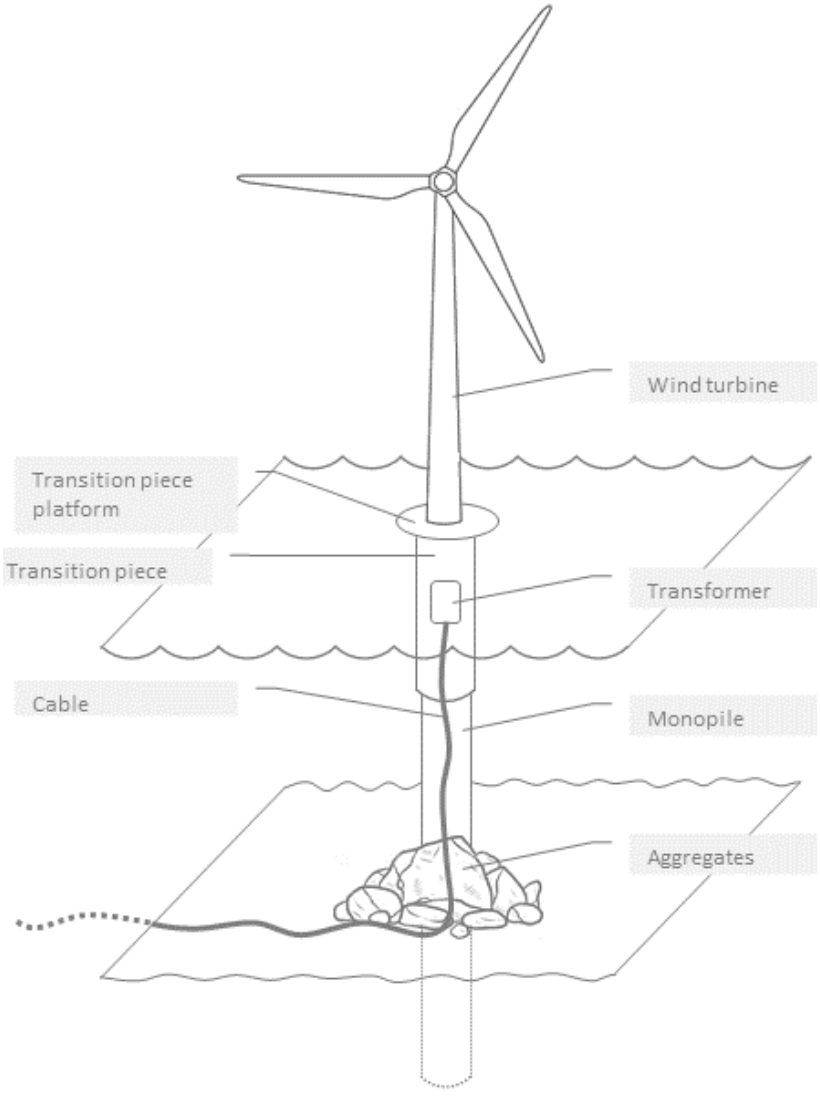

Figure 46-Structure of a monopile foundation for offshore wind turbine[25]

3.5.2 Transformers

Every wind turbine in a wind farm should have a medium voltage transformer to increase the voltage to 33 kV to reduce the transmission loss. When the total capacity of a wind farm is greater than 30 MW, a high voltage transformer will be implemented to increase the voltage to 150 kV and then connected with the national grid [33], these two scenarios are shown in Figure 47.

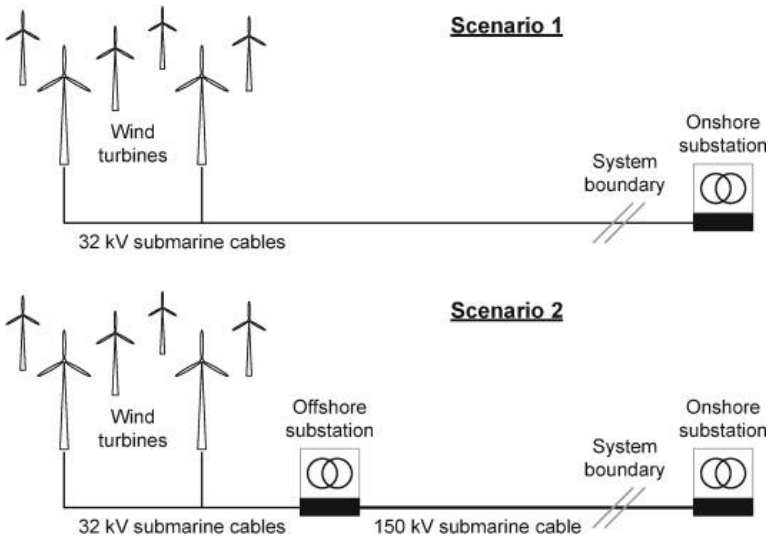

Figure 47-Scenarios for wind farm grid connection[34]

There are available data for transformers with capacities like 315 kVA, 10 MVA, 16 MVA, 40 MVA, 63 MVA, 250 MVA and 500 MVA as shown in Table 23 [35]. Transformers standard ratings are 100, 160, 200, 315, 400, 500, 630, 1000, 1250, 1600, 2000, 2500 kVA for 33 kV [36]. The capacity of transformer is determined by wind turbine rated power, and the size of transformer is assumed as the nominal power of the wind turbine and then used to find the corresponding standard one, although it may lead to an oversize [37], it is a good way to estimate the material inputs and energy supply for transformers. An interpolation is applied to find the total weight of a transformer, and the materials content is calculated by using the average split ratios. Additionally, the electricity and heat demand for assembling is proportional to the transformer size.

It's important to notice that fugitive emissions of sulfur hexafluoride have not been included as their impact is extremely small and the leakage rate has been reduced over time [25].

Table 23- Materials breakdown for transformers [35]

Capacity (MVA)	0.315	10	16	40	63	250	500	Average
Total weight (kg)	1477	27292	40983	62778	81500	197127.	290868	
Electric steel (%)	0.361	0.250	0.254	0.319	0.260	0.341	0.343	0.304
Construction steel (%)	0.219	0.332	0.244	0.253	0.178	0.231	0.184	0.235
Transformer oil (%)	0.230	0.248	0.249	0.247	0.245	0.243	0.217	0.240
Aluminum (%)	0.135	0.002	0.002	0.001	0.000	0.010	0.000	0.022
Insulation material (%)	0.041	0.012	0.016	0.032	0.023	0.041	0.022	0.027
Porcelain (%)	0.007	0.002	0.003	0.003	0.004	0.010	0.009	0.005
Other (%)	0.006	0.003	0.002	0.000	0.038	0.000	0.029	0.011
Copper (%)	-	0.129	0.212	0.145	0.225	0.123	0.137	0.162
Paint (%)	-	0.007	0.005	0.001	0.003	0.000	0.008	0.004
Wood (%)	-	0.013	0.013	-	0.023	-	0.052	0.025

3.5.3 Cables

The length of cables depends on the wind farm layout, location, and total capacity. Offshore wind turbines have a longer distance for electricity transmission (from wind turbine to coast), the distance between wind turbines are defined by rotor diameter, usually 4R between wind turbines, 7R between rows (weak effect). Position of central transformer and wind farm layout differs from project to project, if there is a high voltage transformer, the distance from central transformer to grid is assumed as 20 kilometers based on several reports [38-41].

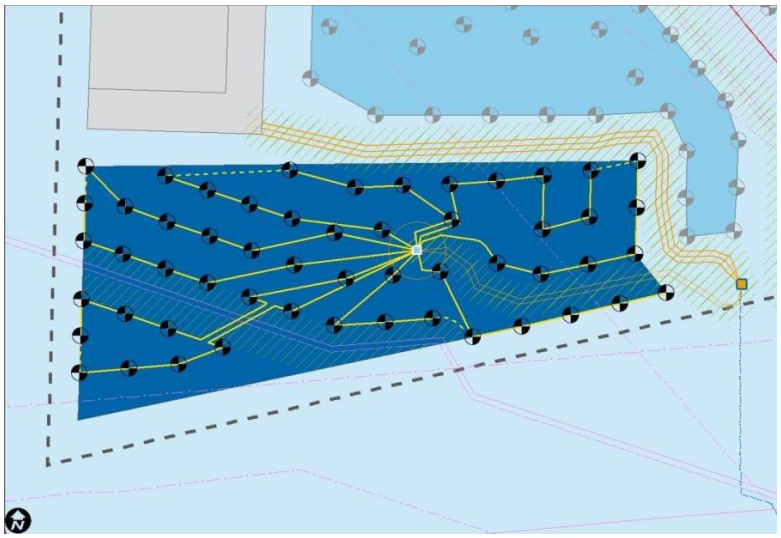

Figure 48-nter-array cabling for a wind farm

The conductor of cables is assumed as copper, and inter-array cabling strategy is applied as shown in Figure 48. The cross section of cables is determined by the voltage and its transport capacity, the higher the capacity, the bigger the cross section. The insulation layer is made from HDPE, PP and PVC, proportional to copper mass [42]. The properties of Copper-based cables can be found from Table 24.

Table 24-Properties of Copper-based cables [43,44]

33 kV cables		150 kV cables	
Cross section (mm^2)	Rated current (A)	Cross section (mm^2)	Rated current (A)
95	352	2000	1560
120	399	1600	1425
150	446	1200	1335
185	502	1000	1160
240	581	800	1045
300	652	630	925
400	726	500	815
500	811	400	700
630	904		
800	993		

It is important to note that the estimation of copper cables may differ from other models for the following reasons, first, the layout of wind farms is strongly depend on the local terrain conditions thus the distances between turbines and the location of central transformer may vary. Second, aluminum-based cables are also used for electricity transmission instead of copper. For offshore wind turbines, the activity of laying cables under seabed is not considered as there is no available data.

3.5.4 Materials breakdown of wind turbine

There isn't enough data to train a model and predict the materials content of a wind turbine for now. An alternative solution is to use some detailed life cycle inventory of wind turbines with different capacities. The following Table 25 contains the detailed inventory for per component of wind turbine as well as the assembly operations for 5 wind turbines with different capacities [42]. This gives a possibility to break down the materials for all the components (rotor, nacelle, tower, foundation, and cables) but with different methods.

Table 25-Overall materials content of wind turbines excluding foundation [42]

Capacity [kW]	30	150	600	800	2000	Average
Steel [%]	0.918	0.813	0.797	0.837	0.725	0.818
Iron [%]	0.040	0.056	0.098	0.062	0.100	0.071
Copper [%]	0.003	0.004	0.003	0.002	0.004	0.003
Aluminum [%]	0.002	0.003	0.003	0.002	0.003	0.003
Polymer [%]	0.017	0.047	0.037	0.037	0.058	0.039
Glass [%]	0.020	0.078	0.062	0.060	0.110	0.066

As shown in the Table 25, the materials split ratio differs from capacities, an average value might be a solution, but not for all cases. Specifically, average values are suitable for Copper, Aluminum as the numbers are extremely close to each other, when it comes to Steel, Iron, the actual values have gaps with averages, which is not negligible, therefore, different methods should be implemented for different case.

If there is an obvious difference in split ratio for materials, interpolation would be applied, like input materials for rotor, there are some big gaps, especially when the capacity is 150 kW, the ratio of fiberglass is higher than 70%, while 40% for 30 kW and 42% for 600 kW, an interpolation based on the wind turbine capacity would be reasonable to improve the result. When the split ratio differences are small for all the capacities, an average value would be calculated for all wind turbines regardless of capacity.

Items like amount of lubricating oil, diesel and electricity demand are calculated based on a function of rated power, simple regression is applied for these items. Table 26 shows a part of detailed inventories of wind turbines with different capacities, and the full inventories can be found from the support information.

Table 26- Inventory for wind turbines with different capacity [42]

		Wind turbine capacity [kW]					Database	Activity name
		30	150	600	800	2000		
Rotor Input	Fiberglass	0.402	0.706	0.421	0.571	0.571	ECOINVENT 3.6 cutoff	market for glass fibre reinforced plastic, polyamide, injection moulded
	Chromium steel	0.380	0.187	0.279	0.211	0.211	ECOINVENT 3.6 cutoff	market for steel, chromium steel 18/8
	Cast iron	0.218	0.107	0.300	0.218	0.218	ECOINVENT 3.6 cutoff	market for cast iron
Rotor Assembly	Chromium steel sheet rolling	0.380	0.187	0.279	0.211	0.211	ECOINVENT 3.6 cutoff	market for sheet rolling, chromium steel
	Steel sheet rolling	0.218	0.107	0.300	0.218	0.218	ECOINVENT 3.6 cutoff	market for sheet rolling, steel
Rotor Disposal	Iron	0.218	0.107	0.300	0.218	0.218	ECOINVENT 3.6 cutoff	market for iron scrap, unsorted
	Steel	0.380	0.187	0.279	0.211	0.211	ECOINVENT 3.6 cutoff	market for scrap steel
	Polymer	0.140	0.247	0.147	0.200	0.200	ECOINVENT 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
	Glass	0.262	0.459	0.273	0.371	0.372	ECOINVENT 3.6 cutoff	treatment of inert waste, inert material landfill
Nacelle Input	Low-alloy steel	0.053	0.053	0.124	0.182	0.182	ECOINVENT 3.6 cutoff	market for steel, low-alloyed
	Chromium steel	0.643	0.643	0.649	0.566	0.493	ECOINVENT 3.6 cutoff	market for steel, chromium steel 18/8
	Cast iron	0.203	0.203	0.143	0.162	0.164	ECOINVENT 3.6 cutoff	market for cast iron
	Rubber	0.003	0.003	0.005	0.005	0.001	ECOINVENT 3.6 cutoff	market for synthetic rubber
	Aluminum	0.016	0.016	0.010	0.010	0.010	ECOINVENT 3.6 cutoff	market for aluminium, wrought alloy
	Copper	0.018	0.018	0.011	0.012	0.012	ECOINVENT 3.6 cutoff	market for copper
	Fiberglass	0.064	0.064	0.059	0.062	0.137	ECOINVENT 3.6 cutoff	market for glass fibre reinforced plastic, polyamide, injection moulded
	Lubricating oil	0.011	0.013	0.002	0.003	0.002	ECOINVENT 3.6 cutoff	market for lubricating oil

*Detailed inventory table can be found from the support information

3.5.5 Transport

Transport of components is strongly dependent on the location of wind turbine and transportation facilities and it may vary from case to case, thus a general proposal is taken from a life cycle assessment report [40] as shown in Table 27.

Table 27-Transport of components to wind plant site [40]

Component	Truck (km)	Ship (km)
Nacelle	1025	0
Hub	1025	0
Blades	600	0
Tower	1100	8050
Foundation	50	0
Other parts	600	0

For maintenance, transportation of crew to and from the site is estimated to be 2160km per plant per year [40].

3.5.6 End of life disposal

For end of life disposal or recycling, associated transport assumed to be 200km to a regional disposal operator, except for foundation, 50km is assumed for waste concrete materials [40].

There are also different strategies for disposals, and again a general proposal is taken from a report[40] as shown in Table 28.

Table 28-End-of-life disposal for different materials [40]

Material	Treatment
Steel	92% recycled, 8% landfilled
Aluminium	92% recycled, 8% landfilled
Copper	92% recycled, 8% landfilled
Polymers	50% incinerated, 50% landfilled
Lubricants	100% incinerated
Others (including concrete)	100% landfilled

3.6 Results and discussions

Different machine learning algorithms are applied for estimating the LCI of wind turbine and also for total electricity production. In most cases, multiple linear regression and random forest regression give better results, polynomial processing and feature extraction will create new features and improve models' performance, and feature selection reduces the dimensionality and simplifies the model. Random forest and neural network are particularly useful for samples with a few features, while support vector machines turn out to be over fitting as the training score is usually high but the validation score is low.

By applying the machine learning models trained for unknown properties and other methods to estimate the components without enough data to building a machine learning model, now it's possible to generate a detailed life cycle inventory for wind turbines even with very limited information.

The Table 29 shows some comparisons between the estimated inventories and materials content from LCA reports [38-41]. For the materials contents, there are some differences but not that huge, however, for the Global Warming Potential, there is a large gap.

The amount of all materials is proportional to the nominal power, which is the same tendency for actual and predicted values, however, the total mass of a wind turbine (including foundation) is somehow over estimated, this is mainly because of the over prediction of foundation, whose main materials are reinforcing steel and concrete, this is why the predicted values of steel and concrete are larger than actual ones, so the model to get the mass of foundation should be modified based on more data. Fiberglass is over predicted, the reason for this might be that the assumption regarding materials breakdown for rotor is very different from these examples. Fiberglass mainly comes from rotor and nacelle, and there is a compromise between fiberglass and cast iron, overestimation of fiberglass leads to decrease of cast iron. Cables for electricity transmission is assumed made of Copper and polymer insulation layers, but actually both Copper and Aluminum (more for economical reason and lower density) are used as conductors, this is why there is a under estimation of Aluminum, another factor is the transmission distance, which strongly depends on the layout of wind farms.

It is also important to notice that even wind turbines with the same manufacturer and capacity may have different contents of materials as shown in Table 30, which makes the estimated life cycle inventory less accurate.

Table 29-Comparison between estimated inventories and materials content from LCA reports

Capacity [kW]	1650[37]		2000[38]		3000[39]		3450[40]	
	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
Offshore	No		No		No		No	
Location	/		/		/		Europe	
Rotor diameter [m]	80		110		90		112	
Hub height [m]	80		80			94.0	94	
Tower weight [ton]	136			152.0		178.0		228.7
Rotor weight [ton]	42.2			56.4		59.4		72.0
Nacelle weight [ton]	51			72.0		98.0		108.9
Foundation weight [ton]	832							
Cables [ton]	15.72							
Transformer [ton]	0.96							
Sum	1077.88							
Steel [ton]	186.40	228.02	234.4	273.96	248.03	332.12	402.52	404.04
Aluminium [ton]	8.68	0.63	10.16	0.87	9.70	1.19	10.21	1.33
Polymer [ton]	18.01	11.15	27.24	13.76	15.30	14.09	29.62	15.08
Copper [ton]	4.51	6.83	3.48	8.71	6.73	9.76	5.03	10.55
Oil [ton]	1.51	1.66	1.48	4.08	1.83	4.35	2.31	4.73
Cast iron [ton]	29.3	17.53	26.56	24.09	32.63	29.02	70.10	33.55
Fiberglass [ton]	8.1	29.99	15.52	42.06	12.37	47.36	26.21	56.05
Electronics [ton]	2.5	0.28	1.84	1.76	1.93	1.50	3.28	1.56
Concrete [ton]	805	823	913.44	972.40	986.67	1223.20	1395.07	1420.32
Wood [ton]	12.63	0.09	/	0.17	/	0.20	/	0.22
Sum	1076.64	1119.38	1234.12	1341.86	1315.20	1662.78	1944.34	1947.44
Steel [%]	0.173	0.204	0.190	0.204	0.189	0.200	0.207	0.207
Aluminium [%]	0.008	0.001	0.008	0.001	0.007	0.001	0.005	0.001
Polymer [%]	0.017	0.010	0.022	0.010	0.012	0.008	0.015	0.008
Copper [%]	0.004	0.006	0.003	0.006	0.005	0.006	0.003	0.005
Oil [%]	0.001	0.001	0.001	0.003	0.001	0.003	0.001	0.002
Cast iron [%]	0.027	0.016	0.022	0.018	0.025	0.017	0.036	0.017
Fiberglass [%]	0.008	0.027	0.013	0.031	0.009	0.028	0.013	0.029
Electronics [%]	0.002	0.000	0.001	0.001	0.001	0.001	0.002	0.001
Concrete [%]	0.748	0.735	0.740	0.725	0.750	0.736	0.718	0.729
Wood [%]	0.012	0.000	/	0.000	0.000	0.000	/	0.000
Sum	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GWP [kg-CO2-e] (100 years per kWh)	7.03E-03	0.0429	7.20E-03	0.0603	6.20E-03	0.0632	5.30E-03	0.0667

Table 30-Materials content for wind turbines with the same capacity [39,41,45,46]

Type of wind turbine	V110	V80	V112	V105
Capacity [kW]	2000	2000	3450	3450
Offshore	No	No	No	No
Location	-	-	Europe	Europe
Rotor diameter [m]	110	80	112	105
Hub height [m]	80	-	94	72.5
Steel [ton]	234.4	256.16	402.52	304.24
Aluminium [ton]	10.16	26.4	10.21	10.21
Polymer [ton]	27.24	49.08	29.62	29.62
Copper [ton]	3.48	6.52	5.03	5.07
Oil [ton]	1.48	1.64	2.31	2.31
Cast iron [ton]	26.56	40	70.10	70.10
Fiberglass [ton]	15.52	19.36	26.21	25.86
Electronics [ton]	1.84	2.4	3.28	3.07
Concrete [ton]	913.44	1105.32	1395.07	1031.62
Wood [ton]	-	-	-	-
sum	1234.12	1506.88	1944.34	1482.10

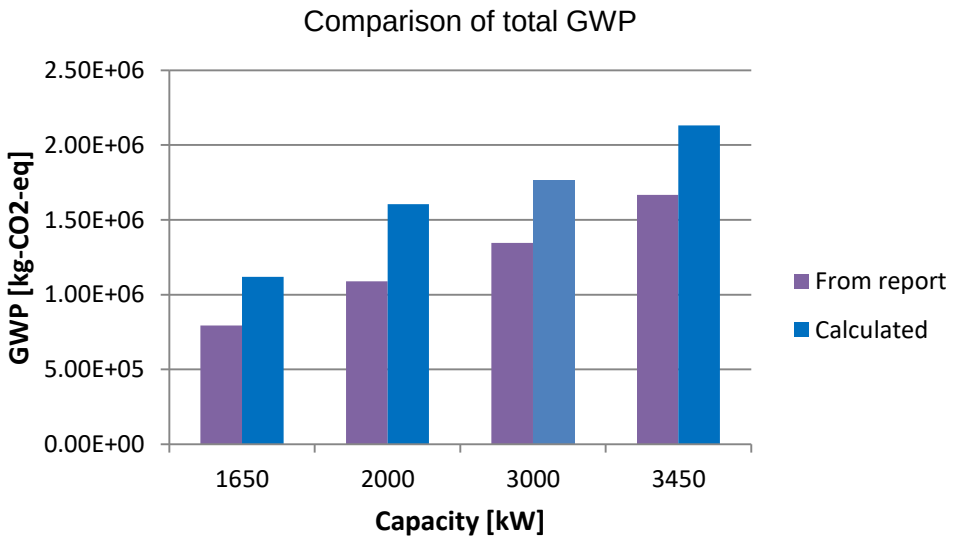

Figure 49-Comparison between GWP calculations and values from report

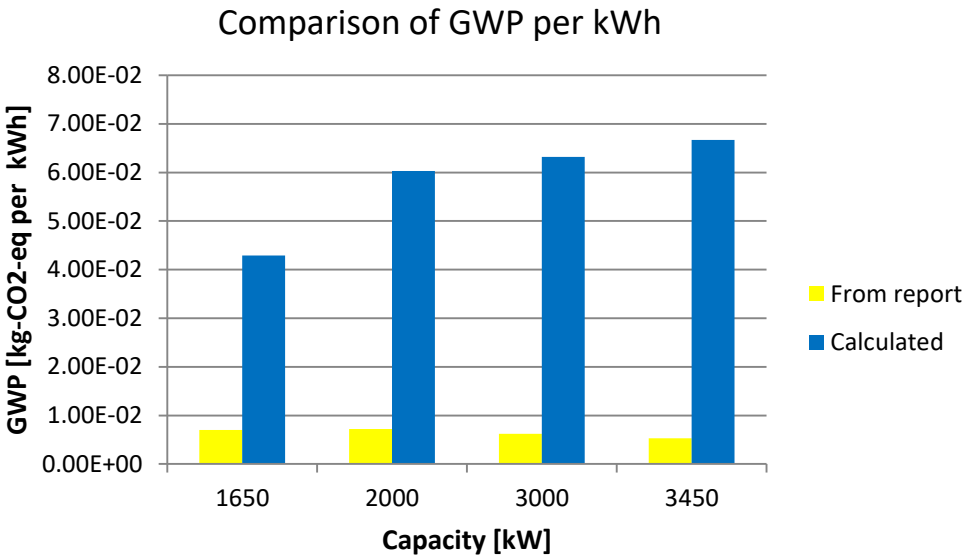

Figure 50-Comparison between GWP/kWh calculations and values from report

When it comes to impact assessment, taking global warming potential as an example, there is a large gap between calculations and the values from the report as shown in Figure 49, the reason for this is the difference of input materials caused by estimation, however, the gaps for GWP per kWh electricity produced are even larger relatively as indicated in Figure 50, apart from differences between input materials, there are some others issues, first, life time of wind turbine and total electricity production, in the reports[38-41], the life time is 20 years and the electricity production is calculated based an annual average load factor over 40%, which may lead to the total electricity production is much more than the actual and predicted values, thus the indexes like GWP/kWh might be smaller. Second, the background processes, since the location of wind farm is either barely mentioned or in a more general way, it's not possible to adjust the background processes to match the location, therefore, datasets from a global level is used as first choices, which may also have negative effect on the impact assessment.

Chapter 4

Conclusions

This chapter summarizes the conclusions from this work including LCI database clustering and LCI estimation of wind turbines. It also points out the aspects to be developed in future works like estimating the entire LCI by machine learning models if there is enough available data, and the methodology of LCI estimation for wind turbines can also be applied to other RES projects.

LCI databases like ECOINVENT and EXIOBASE are important for providing background activities to perform LCA studies, however, no effort has been devoted to explore data structure or other hidden information in the LCI database itself. Clustering is unsupervised machine learning, aims to pattern recognition for unlabeled dataset, there are many different clustering algorithms like partitional clustering, density-based separation and hierarchical clustering with both advantages and drawbacks.

The curse of dimensionality stands for various phenomena that cause difficulties when analyzing or processing data in high-dimensional spaces, for a clustering problem, the dimensionality curse exists as well, fortunately many techniques have been developed to reduce the dimensions of the input dataset. When setting the inventories as input features for LCI database clustering, it's definitely a high-dimensional clustering case as the number of features exceed a thousand, mathematical methods like PCA and correlation analysis are used to reduce the dimensions and retain as much information as possible, besides, engineering solutions like combining similar elementary flows or removing less important flows can help as well.

With many dimensionality reduction techniques, different input features and clustering algorithms being applied, no hidden patterns or other useful information have been founded yet. However, it turns out many LCIA methods are highly correlated, so it's possible to perform LCA in certain impact categories and use them as references for other impact assessments. When adding location of activities as a new feature, the data structure changes significantly even though they still remain as one huge cluster with some outliers, which means if more proper new features are used, there might be better clustering results.

Another issue concerning LCI is that the LCI requires very detailed information from cradle to grave, but in most case they are not available. Then here is another chance to use machine learning algorithms to predict missing data based on limited information, which is known as a supervised machine learning problem, labeled data are being used to train the model and then applied to estimate the unknown data with some initial information. There are many regression models can be applied, Linear Regression is simple and easy to use, Multiple Linear Regression with polynomial features can improve its performance. Random Forest consists of a set of if-then-else decision rules, which usually gives good results with a few input variables. Neural Network includes input layer, output layer and some hidden layers, it is more for classification problems, for regressions, it performs better with fewer input variables but the accuracy is low. Support Vector Machine is also more for classification problems, for

regressions, it tends to be over fitting as the training score is usually high but the testing score is much lower.

There is a tradeoff between bias and variances, to avoid under-fitting or over-fitting, the dataset is usually divided randomly into a training set and a smaller testing set. Model selection can be based on learning curves or some performance indexes. Generally, models with higher testing score and lower errors will be selected for prediction.

In the LCI estimation of wind turbines, a model combined machine learning algorithms and mathematical methods is developed and can be tailored based on the known information. The worst case is that only rated power and location (Onshore or Offshore) of a wind turbine are available, then the dimensions like rotor diameter, hub height, masses of rotor, nacelle, tower can be estimated by machine learning models, sizing of foundation, transformers and required cables are calculated based some mathematical methods and assumptions. Materials breakdown is based on split ratios from some detailed inventories of wind turbines with different capacities, transport requirement, maintenance and end-of-lie disposal strategy is taken from a LCA report.

After estimating the LCI of a wind turbine, it's possible to perform LCA to see the environmental impacts, however, for a better comparison with other projects, it's necessary to know the service time and lifetime electricity production, with available information from Danish Ministry of Energy, it's possible to train models to forecast the life time and total electricity production of wind turbines. Although the model for service time doesn't give satisfying result, it's still better than just setting a life time of 20 years. Furthermore, predicting total electricity without life time is still feasible with a lower accuracy.

The estimated LCI may differ from the actual values, because the materials content for wind turbines vary from project to project and manufacturer to manufacturer. Even some wind turbines with the same capacity and same manufacturer, they may have different input materials, which will make the estimations less accurate.

With more available data, it's achievable to build a model for estimating the entire LCI of a wind turbine by machine learning algorithms. Although the estimated LCI may not be highly accurate, it would give more support for environmental impact assessment and provide references for better decision making especially when the initial information is limited.

This methodology can also be applied for other renewable energy systems like photovoltaic panels and geothermal power plants, in which most of the energy, material requirements and environmental

impacts occur during the manufacturing and installation phases. The first step is to collect data regarding these projects as detailed as possible, next machine learning algorithms can be applied to train models to estimate the input materials by using available information like peak power of a solar panel and its efficiency, as for the activities involved in assembling, maintenance, transport, and end-of-life disposal, they can quantified by other assumptions or mathematical methods, after the estimation of detailed LCI, finally it is possible to implement LCA studies, meanwhile the total electricity production can also be predicted by machine learning models if there is enough information. In the whole process, machine learning is indeed a key role, while LCI database like ECOINVENT providing the background information is essential as well, because it is the fundamental of this methodology and makes it feasible to perform impact assessment by only quantifying the involved activities.

References

[1] Mark A. J. Huijbregts, Linda J. A. Rombouts, Stefanie Hellweg, Rolf Frischknecht, A. Jan Hendriks, Dik van de Meent, Ad M. J. Ragas, Lucas Reijnders, and Jaap Struijs. Is cumulative fossil energy demand a useful indicator for the environmental performance of products? Environmental Science & Technology (2006): 641-648.

[2] Esnouf A, Heijungs R, Coste G, Latrille É, Steyer J.P, and Hélias A. A tool to guide the selection of impact categories for LCA studies by using the representativeness index. Science of the Total Environment, 2019, 658: 768-776.

[3] Steinmann Z J N, Schipper A M, Hauck M, Giljum S, Wernet G, and Huijbregts M.A. Resource footprints are good proxies of environmental damage. Environmental science & technology, 2017, 51(11): 6360-6366.

[4] Berger M, Finkbeiner M. Correlation analysis of life cycle impact assessment indicators measuring resource use. The International Journal of Life Cycle Assessment, 2011, 16(1): 74-81.

[5] Wernet G, Bauer C, Steubing B, Reinhard J, Moreno-Ruiz E, and Weidema B. The ECOINVENT database version 3 (part I): overview and methodology. The International Journal of Life Cycle Assessment, 2016, 21(9): 1218-1230.

[6] Merciai, S. and J. Schmidt. Methodology for the Construction of Global Multi-Regional Hybrid Supply and Use Tables for the EXIOBASE v3 Database. Journal of Industrial Ecology, 2018, 22(3): 516-531.

[7] Mutel C. Brightway: an open source framework for life cycle assessment. Journal of Open Source Software, 2017, 2(12): 236.

[8] Kluyver T, Ragan-Kelley B, Pérez F, Granger B.E, Bussonnier M, Frederic J, and Ivanov P. Jupyter Notebooks-a publishing format for reproducible computational workflows. ELPUB. 2016: 87-90.

[9] Bishop C M. Pattern recognition and machine learning. springer, 2006.

[10] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. Stanford, 2006.

[11] Grira N, Crucianu M, Boujemaa N. Unsupervised and semi-supervised clustering: a brief survey. A review of machine learning techniques for processing multimedia content, 2004, 1: 9-16.

[12] Ester M, Kriegel H P, Sander J, and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996, 96(34): 226-231.

[13] Ankerst M, Breunig M M, Kriegel H P, Ng R.T, and Sander J. OPTICS: ordering points to identify the clustering structure. ACM Sigmod record, 1999, 28(2): 49-60.

[14] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Record, 1996, 25(2): 103-114.

[15] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(5): 603-619.

[16] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 1987, 20: 53-65.

[17] Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 1974, 3(1): 1-27.

[18] Davies D L, Bouldin D W. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, 1979 (2): 224-227.

[19] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering In: Journal of Open Source Software, The Open Journal, volume 2, number 11. 2017

[20] Vidal R. Subspace clustering. IEEE Signal Processing Magazine, 2011, 28(2): 52-68.

[21] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.

[22] Bryant F B, Yarnold P R. Principal-components analysis and exploratory and confirmatory factor analysis. 1995.

[23] Schölkopf B, Smola A, Müller K R. Kernel principal component analysis. International conference on artificial neural networks. Springer, Berlin, Heidelberg, 1997: 583-588.

[24] Tenenbaum J B, De Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction. science, 2000, 290(5500): 2319-2323.

[25] Sacchi R, Besseau R, Perez-Lopez P, and Blanc I. Exploring technologically, temporally and geographically-sensitive life cycle inventories for wind turbines: A parameterized model for Denmark. Renewable Energy, 2019, 132: 1238-1250.

[26] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, and Vanderplas J. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 2011, 12: 2825-2830.

[27] Evgeniou T, Pontil M. Support vector machines: Theory and applications. Advanced Course on Artificial Intelligence. Springer, Berlin, Heidelberg, 1999: 249-257.

[28] Energinet.dk, Stamdataregister for Vindkraftanlæg, 2017. https://ens.dk/service/statistik-data-noegletal-og-kort/data-oversigt-over-energisektoren.

[29] Staffell, Iain and Pfenninger, Stefan. Using Bias-Corrected Reanalysis to Simulate Current and Future Wind Power Output. 2016, Energy 114, pp. 1224-1239.

[30] Rui Castro, Wind Power, Chapter 4, Renewable Energy Sources and Dispersed Power Generation, Edition 0, October 2018.

[31] Tall towers for large wind turbines, Report from Vindforsk project V-342 Höga torn för vindkraftverk, Elforsk rapport 10:48

[32] Energinet.dk, Technical Project Description for Offshore Wind Farms (200MW), 2015.

[33] Nexans, Integrated Cable Solutions for Offshore Wind Development, 2015.

[34] Huang Y F, Gan X J, Chiueh P T. Life cycle assessment and net energy analysis of offshore wind power systems. Renewable Energy, 2017, 102: 98-106.

[35] Jorge R S, Hawkins T R, Hertwich E G. Life cycle assessment of electricity transmission and distribution—part 2: transformers and substation equipment. The International Journal of Life Cycle Assessment, 2012, 17(2): 184-191.

[36] 3-Phase Distribution Transformers 11 or 33 kV/415-240V. https://www.mstcecommerce.com/auctionhome/RenderFileViewVideo.jsp?file=ddugjy-3-Phase-DTs.pdf

[37] De Caro S, Scimone T, Testa A, and La Torre R. Optimal size selection for step-up transformers for wind generation plants. International Symposium on Power Electronics Power Electronics, Electrical Drives, Automation and Motion. IEEE, 2012: 571-576.

[38] Life Cycle Assessment of Electricity Production from an onshore power plant based on Vestas V82-1.65 MW turbines. Vestas Wind Systems A/S, 2006.

[39] Razdan P, Garrett P. Life Cycle Assessment of Electricity Production from an onshore V110-2.0 MW Wind Plant. Vestas Wind Systems A/S, 2015.

[40] Garrett P, Ronde K. Life Cycle Assessment of Electricity Production from an onshore V90-3.0 MW Wind Plant. Vestas Wind Systems A/S, 2013.

[41] Razdan P, Garrett P. Life Cycle Assessment of Electricity Production from an onshore V112-3.45 MW Wind Plant. Vestas Wind Systems A/S, 2017.

[42] B. Burger, C. Bauer, Teil XIII - Windkraft, (2007). http://windland.ch/doku_wind/06_XIII_Windkraft.pdf.

[43] Nexans, XLPE Insulated Cable 150 KV, 2017.

[44] Nexans, 2XS(FL)2YRAA RM 19/33 (36)kV, 2016.

[45] Garrett P, Ronde K. Life Cycle Assessment of Electricity Production from a V80-2.0 MW Gridstreamer Wind Plant. Vestas Wind Systems A/S, 2011.

[46] Razdan P, Garrett P. Life Cycle Assessment of Electricity Production from an onshore V105-3.45 MW Wind Plant. Vestas Wind Systems A/S, 2017.

Support information

The following Table 31 gives detailed inventories for wind turbines with different capacities, including 30 kW, 150 kW, 600 kW, 800 kW, and 2000 kW. All the input materials are calculated based on the split ratio in these inventories by interpolation, average or simple regression.

Table 31- Detailed inventory for wind turbines with different capacity[42]

		Wind turbine capacity [kW]					Database	Activity name
		30	150	600	800	2000		
Rotor Input	Fiberglass	0.402	0.706	0.421	0.571	0.571	ecoinvent 3.6 cutoff	market for glass fibre reinforced plastic, polyamide, injection moulded
	Chromium steel	0.380	0.187	0.279	0.211	0.211	ecoinvent 3.6 cutoff	market for steel, chromium steel 18/8
	Cast iron	0.218	0.107	0.300	0.218	0.218	ecoinvent 3.6 cutoff	market for cast iron
Rotor Assembly	Chromium steel sheet rolling	0.380	0.187	0.279	0.211	0.211	ecoinvent 3.6 cutoff	market for sheet rolling, chromium steel
	Steel sheet rolling	0.218	0.107	0.300	0.218	0.218	ecoinvent 3.6 cutoff	market for sheet rolling, steel
Rotor Disposal	Iron	0.218	0.107	0.300	0.218	0.218	ecoinvent 3.6 cutoff	market for iron scrap, unsorted
	Steel	0.380	0.187	0.279	0.211	0.211	ecoinvent 3.6 cutoff	market for scrap steel
	Polymer	0.140	0.247	0.147	0.200	0.200	ecoinvent 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
	Glass	0.262	0.459	0.273	0.371	0.372	ecoinvent 3.6 cutoff	treatment of inert waste, inert material landfill
Nacelle Input	Low-alloy steel	0.053	0.053	0.124	0.182	0.182	ecoinvent 3.6 cutoff	market for steel, low-alloyed
	Chromium steel	0.643	0.643	0.649	0.566	0.493	ecoinvent 3.6 cutoff	market for steel, chromium steel 18/8
	Cast iron	0.203	0.203	0.143	0.162	0.164	ecoinvent 3.6 cutoff	market for cast iron
	Rubber	0.003	0.003	0.005	0.005	0.001	ecoinvent 3.6 cutoff	market for synthetic rubber
	Aluminum	0.016	0.016	0.010	0.010	0.010	ecoinvent 3.6 cutoff	market for aluminium, wrought alloy
	Copper	0.018	0.018	0.011	0.012	0.012	ecoinvent 3.6 cutoff	market for copper
	Fiberglass	0.064	0.064	0.059	0.062	0.137	ecoinvent 3.6 cutoff	market for glass fibre reinforced plastic, polyamide, injection moulded
	Lubricating oil	0.011	0.013	0.002	0.003	0.002	ecoinvent 3.6 cutoff	market for lubricating oil

		Wind turbine capacity [kW]					Database	Activity name
		30	150	600	800	2000		
Nacelle Assembly	Aluminum sheet rolling	0.016	0.016	0.010	0.010	0.010	ecoinvent 3.6 cutoff	market for sheet rolling, aluminum
	Chromium steel sheet rolling	0.643	0.643	0.649	0.566	0.493	ecoinvent 3.6 cutoff	market for sheet rolling, chromium steel
	Steel sheet rolling	0.256	0.256	0.267	0.345	0.347	ecoinvent 3.6 cutoff	market for sheet rolling, steel
	Electricity [kWh]	575*	3987	17510	17510	67500	ecoinvent 3.6 cutoff	market for electricity, medium voltage
Nacelle Disposal	Steel	0.695	0.695	0.772	0.748	0.675	ecoinvent 3.6 cutoff	market for scrap steel
	Iron	0.203	0.203	0.143	0.162	0.164	ecoinvent 3.6 cutoff	market for iron scrap, unsorted
	Aluminum	0.016	0.016	0.010	0.010	0.010	ecoinvent 3.6 cutoff	market for scrap aluminium
	Copper	0.018	0.018	0.011	0.012	0.012	ecoinvent 3.6 cutoff	market for scrap copper
	Used oil	0.011	0.013	0.002	0.003	0.002	ecoinvent 3.6 cutoff	market for waste mineral oil
	Polymer	0.026	0.026	0.021	0.027	0.040	ecoinvent 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
	Glass	0.042	0.042	0.038	0.041	0.097	ecoinvent 3.6 cutoff	treatment of inert waste, inert material landfill
Tower Input	Low-alloy steel	0.993	0.994	0.996	0.995	0.995	ecoinvent 3.6 cutoff	market for steel, low-alloyed
	Epoxy resin	0.007	0.006	0.004	0.005	0.005	ecoinvent 3.6 cutoff	market for epoxy resin, liquid
Tower Assembly	Steel arc welding [m]	84	115	152	190	228	ecoinvent 3.6 cutoff	market for welding, arc, steel
	Galvanizing [m^2]	74	190	625	818	1978	ecoinvent 3.6 cutoff	market for zinc coat, pieces
	Steel sheet rolling	0.993	0.994	0.996	0.995	0.995	ecoinvent 3.6 cutoff	market for sheet rolling, steel
Tower Disposal	Low-alloy steel	0.993	0.994	0.996	0.995	0.995	ecoinvent 3.6 cutoff	market for scrap steel
	Epoxy resin	0.007	0.006	0.004	0.005	0.005	ecoinvent 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
Foundation Input	Concrete	0.956	0.956	0.941	0.941	0.949	ecoinvent 3.6 cutoff	market for concrete, sole plate and foundation
	Reinforcing steel	0.044	0.044	0.059	0.059	0.051	ecoinvent 3.6 cutoff	market for reinforcing steel
	Transformation [m^2]	1004	1004	1091	1121	1318	ecoinvent 3.6 cutoff	Transformation, from pasture and meadow
	Conversion (road) [m^2]	1000	1000	1000	1000	1000	ecoinvent 3.6 cutoff	Transformation, to traffic area, road network

* Actual demands are not split ratio, a simple regression will be used to calculate the demand for other capacities.

		Wind turbine capacity [kW]					Database	Activity name
		30	150	600	800	2000		
Foundation Input	Conversion (industrial) [m^2]	4	4	91	121	318	ecoinvent 3.6 cutoff	Transformation, to industrial area, built up
	Use of road [m^2]	1000	4000	40000	40000	40000	ecoinvent 3.6 cutoff	Occupation, traffic area, road network
	Use of industrial area [m^2]	4	160	3640	4840	11800	ecoinvent 3.6 cutoff	Occupation, industrial area, built up
Foundation Assembly	Diesel [MJ]	5381	5381	27360	29640	74500	ecoinvent 3.6 cutoff	market for diesel, burned in building machine
	Explosive [kg]	10	10	10	10	10	ecoinvent 3.6 cutoff	market for explosive, tovex
	Electricity [kWh]	10	10	10	10	10	ecoinvent 3.6 cutoff	market for electricity, medium voltage
Foundation Disposal	Concrete	0.956	0.956	0.941	0.941	0.949	ecoinvent 3.6 cutoff	treatment of inert waste, inert material landfill
	Steel	0.044	0.044	0.059	0.059	0.051	ecoinvent 3.6 cutoff	market for scrap steel
Power Supply Input	Total weight (kg)	617	1174	2068	2259	-	ecoinvent 3.6 cutoff	
	Copper	0.357	0.341	0.511	0.539	-	ecoinvent 3.6 cutoff	market for copper
	HDPE granules	0.355	0.373	0.279	0.263	-	ecoinvent 3.6 cutoff	market for polyethylene, high density, granulate
	PP granules	0.032	0.017	0.010	0.009	-	ecoinvent 3.6 cutoff	market for polypropylene, granulate
	PVC impact resistant	0.256	0.269	0.201	0.189	-	ecoinvent 3.6 cutoff	market for polyvinylchloride, bulk polymerised
Power Supply Assembly	Copper wire drawing	0.357	0.341	0.511	0.539	-	ecoinvent 3.6 cutoff	market for wire drawing, copper
Power Supply Disposal	Copper	0.357	0.341	0.511	0.539	-	ecoinvent 3.6 cutoff	market for scrap copper
	HDPE granules	0.355	0.373	0.279	0.263	-	ecoinvent 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
	PVC impact resistant	0.256	0.269	0.201	0.189	-	ecoinvent 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
	PP granules	0.032	0.017	0.010	0.009	-	ecoinvent 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
Transformer Input	Electric steel						ecoinvent 3.6 cutoff	steel production, electric, low-alloyed
	Construction steel						ecoinvent 3.6 cutoff	market for steel, unalloyed
	Transformer oil						ecoinvent 3.6 cutoff	market for lubricating oil
	Aluminium						ecoinvent 3.6 cutoff	market for aluminium, wrought alloy

		Wind turbine capacity [kW]					Database	Activity name
		30	150	600	800	2000		
Transformer Input	Insulation material						ecoinvent 3.6 cutoff	market for glass wool mat
	Porcelain						ecoinvent 3.6 cutoff	market for ceramic tile
	Copper						ecoinvent 3.6 cutoff	market for copper
	Paint						ecoinvent 3.6 cutoff	market for electrostatic paint
	Wood						ecoinvent 3.6 cutoff	planing, board, softwood, u=20%
Transformer Assembly	Copper wire drawing						ecoinvent 3.6 cutoff	market for wire drawing, copper
	Steel sheet rolling						ecoinvent 3.6 cutoff	market for sheet rolling, steel
	Aluminium sheet rolling						ecoinvent 3.6 cutoff	market for sheet rolling, aluminium
	Electricity [kWh]						ecoinvent 3.6 cutoff	market for electricity, medium voltage
	Heat [kWh}						ecoinvent 3.6 cutoff	heat, from municipal waste incineration to generic market for heat district or industrial, other than natural gas
Transformer Disposal	Copper						ecoinvent 3.6 cutoff	market for scrap copper
	Steel						ecoinvent 3.6 cutoff	market for scrap steel
	Used oil						ecoinvent 3.6 cutoff	market for waste mineral oil
	Aluminium						ecoinvent 3.6 cutoff	market for scrap aluminium
	Polymer						ecoinvent 3.6 cutoff	treatment of waste plastic, mixture, municipal incineration
	Porcelain						ecoinvent 3.6 cutoff	treatment of inert waste, inert material landfill
Maintenance	Lubricating oil	84	168	360	450	1000	ecoinvent 3.6 cutoff	market for lubricating oil
	Car [km]						ecoinvent 3.6 cutoff	market for transport, passenger car

-Empty means no data available from the report

-For transformers, the data is calculated from a model define before

-For maintenance, main purpose is to change the lubricating oil, transport of technicians by car is define in the transport part