Regression of species richness biodiversity indicator from satellite observations and environmental parameters

MALIS Project Final Report, Fall Semester 2020

Vincenzo Madaghiele, Cosimo Chetta

***Abstract*—Measuring biodiversity is often a complex and resource demanding task, which requires observations on the field. The European Union's satellite infrastructure provides a constant stream of data, which is already used for different ecological monitoring applications. This experiment focuses on measuring the species richness biodiversity indicator from satellite observations and environmental parameters only, in order to provide a scalable, habitat-level method for biodiversity measuring. After building an appropriate database by collecting and aggregating different environmental variables, we have focused on comparing the performances of different regression models on the data, then we have experimented with a more complex neural network architecture, including satellite images into the datasets.**

I. Introduction

Ecosystems provide the basic support to human life. They ensure that the climatic conditions on Earth are stable, provide indispensable nutrients to the soil and regulate the air we breathe. The survival of ecosystems depends upon the survival of the biodiversity that they host. The more an ecosystem is biodiverse, the more it is stable and able to cope with changes. Therefore, it is fundamental to protect and monitor the biodiversity on Earth, which is decreasing at a fast pace due to human activities. Biodiversity is difficult to measure because it is defined by a series of very complex relations among species, so many variables have to be taken into account in order to explain biodiversity changes. Our experiment aims at exploiting the data generated by the EU satellite technology and sensor networks to provide a reliable estimation of the conditions of biodiversity in a given region. The code we used to run our experiment can be found at [1].

II. Related work

Measuring biological diversity is a complex task, which has been approached in many different ways in the literature. An excellent summary of the biodiversity metrics definitions and methods of measurement is offered in [2]. In particular in Chapter 4, the book describes the main techniques used to assess species richness. A comparison between the modeling algorithm used for richness estimation is done in [3]. As the importance of having a standardized global biodiversity monitoring rises, ecologists have defined a set of Essential Biodiversity Variables (EBVs), currently in stage of standardization ([4]). In our process we have therefore followed the method described in [5], which defines the stages for data collection and production regarding the Biodiversity Distribution and Abundance data for the EBVs. As highlighted in [6], the use of machine learning techniques for ecological monitoring from remote sensing seems to be in its early stages, and it is important to connect the great amount of data coming from the remote sensing infrastructure with the research field of ecological modeling. In the literature a vast array of machine learning techniques have been used for species richness estimation in different habitats. In particular, random forest has been used for assessing feature importances and to make ecological prediction in [7], [8] for terrestrial ecosystems, and in [9] and [10] for marine ecosystems. More similar to our experiments are [11] and [12], which both approach the problem of forest related species richness from a different angle. [11] compares the effect of different groups of features on the species richness prediction in Switzerland, while [12] compares the importances of different features and the performance of linear models with respect to random forest in the prediction of vascular plant species richness in Chile. A common choice of variables for forest related experiments such as ours includes features derived from high-resolution LiDAR images, hyperspectral images, field data and variables that can be derived from satellite observations. One of the most significant differences between our experiment and the cited studies is that all of the features we have chosen are general-purpose and can be derived from satellite observations. In line with the definition of the EBVs, the aim of this project was to provide a reliable, scalable model for biodiversity assessment, so we have chosen to use features which are continuously collected and provided with open access by the Copernicus program. For example, while in [11] fewer, more specific variables are used, such as Canopy Height and Vegetation Density, our approach was to use a higher number of general-purpose variables, which can all be estimated by remote sensing and do not require field measurements. Another difference between our experiment and the ones cited before is the spatial resolution, which in general tends to be in the range of meters, while in our case the features provided by the Copernicus Program have a precision of 300 meters or 1 km. As for the individual features choice, a comparison of the effect different features could have on the prediction is made in [13], which describes the effect of different groups of variables on the regression of species richness.

III. Dataset and features

A. Dataset builduing

We have built our dataset using publicly available geospatial data offered by the Copernicus EU space program and the European Environment Agency (EEA). This dataset is composed of features from three main sources:

- **Copernicus Land** [14]
- **Copernicus Climate Change** [15]
- **EEA Richness of forest-related species indicator** [16]

As pointed out in [17], time and space scale are fundamental in biodiversity monitoring, and scales have to be chosen according to the specific phenomena to be measured. The aim of our experiment is to provide a reliable estimation of the overall status of biodiversity in a given area, so we have opted for a community level diversity measurement, and we choose time and space scales accordingly. The Land, Climate and Atmosphere datasets were downloaded taking into account data of the same year and month. When the data was not available for the selected year, we have dowloaded it for the same month of the nearest year. We have chosen to do so because of the high seasonal dependence of the data. We have chosen a spatial sampling rate of 0.01 degree of latitude/longitude, so each sample in our dataset contains the data of a cell of 0.01 x 0.01 degrees of latitude/longitude. A summary and a description of the chosen features is available at [18].

B. Selection of biodiversity indicator

Measuring biodiversity is a complex task, and many approaches have been proposed in the literature ([19]), depending on the specific need and application of the measure. It is possible to measure biodiversity on all levels of life, from ecosystem diversity to genetic diversity among the same population. In the case of our experiment, we needed a global indicator which could state the overall status of the biodiversity in an area inside the same habitat, and could also be publicly available in a geospatial dataset for download, so we choose to measure species diversity inside the same region. The indicator we used is the **Richness of forest-related species and habitats indicator**. This index was developed by the European Environment Agency in order to asses the conservation status of European forest habitats. It expresses the richness of species related to forest habitats as a number between 0 and 1, calculated using a series of sub-indicators. It is publicly available just for European territory [16], and just for one time period in 2012. For a more in depth definition of the species richness and the implications of its usage as an indicator we refer to page 39 of [2].

C. Area selection

Building a training set for this experiment requires a prior knowledge of the basic ecology of these areas. We have constructed four basic datasets, which represent different biodiversity hotspots in Europe. We have selected the areas among the Sites of Community Importance (SCI), defined in the European Commission Habitats Directive (92/43/EEC) and in the European Commission Bird Directive (79/409/EEC), protected under the Natura 2000 network. We have considered the different types of forest areas and the different biogeographycal regions. We have also considered the internal ecological similarity of all these data, in order to obtain homogeneous datasets. The chosen areas are shown in table I.

TABLE I: Selected areas

State	*Latitude*	*Longitude*	*Biogeographical region(s)*
France	3.6-4.5	44.4-44.8	Mediterranean, Alpine
Finland	25.0-28.0	67.0-69.0	Boreal, Alpine
Italy	12.5-15.5	40.5-43.0	Continental, Mediterranean, Alpine
Bulgaria	22.0-27.0	41.0-44.0	Continental, Alpine

D. Data cleaning

The Copernicus Climate database values are well defined, there are no missing values and from the exploration of the data we have not found any particular issue therefore we have not performed any computation. It has a precision of 0.1 degree of latitude/longitude so we have joined this dataset with the others by finding the closest point. Since the two remaining datasets have a precision of 0.01 degree, one row of Copernicus Climate has been joined with approximately other ten rows of Land/Richness. The EEA Richness of forest-related species indicator is supposed to have values in range (0,1]. Values less than 0 signal a non-forest areas therefore we have removed those coordinates. For the Copernicus Land database, we have tried different approach to handle the non valid data:

- Remove the rows with invalid features
- Fill with mean of the feature
- Fill with the mean of the four closest coordinates
- Fill with a KNN-imputer

The Copernicus Land documentation reports information about errors in the data acquisition. We have considered the errors in data acquisition as invalid, and we have set the values which were too high or too low respectively to the maximum and the minimum values of the feature. We have considered values that are outside the feature boundaries reported in the documentation as invalid. Finally, by plotting the data of each feature we found some values of the Albedo (ALBH/ALDH) and Top-Of-Canopy Reflectance (TOCR) features that could be considered as outliers. For those two group of features, we set as invalid the values with a z-score greater than 3. After having merged the three dataset over the latitude-longitude coordinates, we tested regional data using a 4 fold cross validation on a random forest regressor.

As can be seen from the results in Table II, the only outliers handler that caused a meaningful score reduction is the removal of rows with outliers. The other techniques have similar scores, so we chose to continue with the mean of closest coordinates since it performed slightly better than the other two.

TABLE II: Outlier handling methods performances

Region	***Remove***	***Mean***	***Closest mean***	***knn imputer***
Bulgaria	0.918	0.925	0.925	0.923
Finland	0.713	0.717	0.717	0.717
France	0.732	0.742	0.742	0.735
Italy	0.803	0.832	0.835	0.827

E. Image data

A wide range of image data is available from ESA's Sentinel program. We have considered using data from Sentinel-1 and Sentinel-2 missions. Sentinel-1 uses a synthetic-aperture radar (C-SAR), which provides measurements in the microwave range (central frequency at 5.405 GHz), while Sentinel-2 data is in the frequency range of the visible light. We have experimented with the two different types of images, and we have found that in most of the Sentinel-2 images available for the areas of our datasets the ground was covered by clouds. We have therefore decided to use Sentinel-1 data, whose frequency range is not influenced by clouds. We have used the Sentinel-1 Level-1 Ground Range Detected (GRD), VV (Partial Dual polarisation) images. We have chosen to experiment with the Italy area of our datasets. We have segmented the images according to latitude and longitude, with 0.01 degree precision (as for the features dataset). Before segmentation it was necessary to georeference the images, because the grd data products are not georeferenced by default. The product of the image segmentation is a single 1 band image in geoTiff format for each line in the corresponding features dataset. The images were then converted to .jpg format and than rescaled to 91x91 size to be used as input to the model.

IV. Methods

A. Regression models

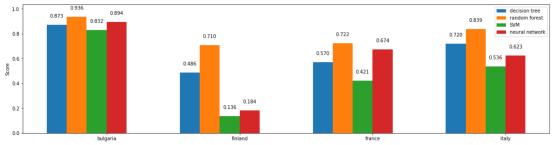
After obtaining the dataset, we have performed various experiments to select the best model. We have compared four different regression techniques: decision tree, random forest, SVM and neural network. We have made a grid search for each model in order to obtain the best performing parameters and validated it using a 4-fold cross validation. In order to measure and compare performances we have used the coefficient of determination R^2 ([20]).

B. Images model architecture

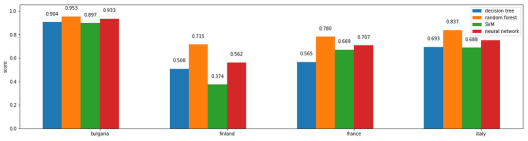
To implement the images in our model, we have used a modified version of the AlexNet [21] to extract the feature from the images, paired it with a Neural Network that worked on the previous data and then combined the two outputs over a third NN to get the regression results. Since it would have been too time consuming to perform a cross-validation, we have divided the dataset using a 70-15-15 train, validation, test split. We have run the model for 200 epochs, performing a validation every 5 epochs and then tested the best one. We have repeated this operation three times to get an average score that could be as accurate as possible.

V. Experimental Results

A. Comparison of regression models



(a) R^2 score of the model for each region for a 4-fold cross validation



(b) R^2 score of the model for each region for a 4-fold cross validation on log(species richness)

Fig. 1: R^2 score of the models for each region

As can be seen from the Figure 1, the performances are different depending on the model and on the dataset. The best results have been obtained using random forest on the Bulgaria dataset, while the worst results regard SVM on the Finland dataset. Comparing the results in Figure 1 it is possible to see that for each geographical area there is a consistent difference between the models; specifically the SVM models perform significantly worse on France and Finland data and Neural Network has particularly low score on Finland. These low results can be explained by investigating the different datasets: by calculating the variance of the species richness label for each dataset we can see that Finland and France have a significantly lower variance than the other datasets. This means that the labels are really close to each other in the feature space and it is difficult for the model to distinguish between them. To solve this problem we have applied a log function to the species richness label, in order to separate them by expanding the feature space. This technique has indeed proved useful, the results of SVM and Neural Network have significantly improved, as can be seen from Figure 1b. From the Figure it is possible to notice that, as expected, the Random Forest and Regression Tree results have been little affected by the application of the log for feature spreading and that the result of Random Forest is still the best one for each model. However, the ratios between the results of each area are now more similar to each other and we can conclude that the difference among the areas are due to the internal composition of the datasets. The Bulgaria dataset better represents the regression index because of the higher internal variability, while the Finland data is more homogeneous and it is not able to correctly describe some of the more different points.

Another consideration should be done on the size and the internal composition of the chosen areas. We have found that the result of our experiment is very dependent on the locality of the dataset, and even though we have trained on fairly large, region-wide areas which comprise fairly different habitats we

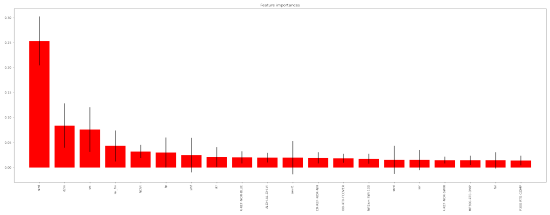
have found that each model has poor performance on habitats which it has never seen even partially in the training set. This means that the accuracy of the prediction is area-dependent and that the model is not able to generalize to vast areas. The prediction is also influenced by the size of the chosen area, in fact we have found that choosing too large or too small areas decreases the prediction score. This could be because the complex relationships that characterize biodiversity on a species level are best described on a medium size area. If the area is too big the feature values become too diverse and lose connection between each other, spanning across ecosystems with different characteristics; the model is therefore not able to connect features that are so diverse and fails. On the other hand, if the area is too small and too internally homogeneous, the model is not able to learn the general meaning of each feature and it struggles to classify new points. The extent to which this characteristic influences the model could of course be depending on the spatial resolution of our dataset. This result was somehow expected, because it confirms the different role that each feature plays in a different forest ecosystem.

It is possible to observe that the Random Forest model has given the best result with respect to all the other methods for all of the areas, reaching satisfying performance in all of them. This confirms, as said in [11], the capability of Random Forest to capture the non-linearities and the interactions among species in the ecosystem better than other models. Overall, it is worth noticing the relative ratio of R^2 for each method is similar for all areas and this means that, even if the areas are different, there is a coherence among the type of features and how they react to different models.

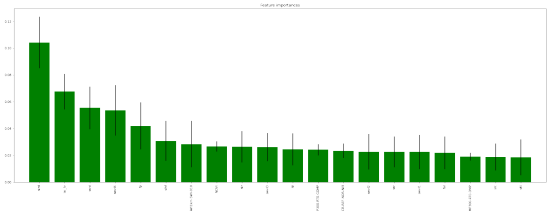
B. Feature importance

We have used random forest to highlight the contribution of each feature to the biodiversity of each area. In order to do so we have trained a different random forest on each dataset and printed the feature importances as a histogram. In figure 2 it is possible to see an example of the importance histogram of France and Finland. The features are displayed on the x axis in order of importance, so it is possible to see how much each feature contributes to the result.

As can be seen from the figure, even though the most important features are similar for the two areas, the importances are distributed in a different way. For example in the graph of Finland many features have a high contribution to the result, while in the case of France the importance is concentrated into fewer features. This could be due to the differences in environmental conditions of these areas, but also to the degree of internal homogeneity of the datasets. More homogeneous areas with fewer, more defined habitats are better described by a smaller set of features, while bigger areas with higher internal variability are more difficult to describe and predict. An in-depth analysis of the importance of each feature and its relative effect on the species richness is beyond the scope of this experiment. However we can observe that most of the prominent variables are coming from the Copernicus Climate Change service (in lowercase in Figure 2)



(a) Most important features for the France dataset (first 20 features)



(b) Most important features for the Finland dataset (first 20 features)

Fig. 2: Feature importances

and in forest related features, such as Leaf Area index (lai) or NDVI. This is highlighting the fact that, as also stated in [11], species richness in forest habitats is predominantly influenced by climatic variables. Of course, collaboration with ecologists would be needed for a better analysis and interpretation of the importance data.

C. Visualization of the regression tree

Random forest and regression trees have been proven useful in many ecological monitoring and forecasting application ([7], [8]), because they provide a good alternative to traditional mathematical models and are able to capture interactions in complex systems such as habitats. Being able to effectively visualize the results of the regression is crucial for the practical applications of this model, because it allows to understand the model's decisions. As said in section V-A, the best result in the regression task has been obtained by the random forest. However, visualizing the result of random forest in its entirety would not give many significant information, because the model is too large to be interpreted. In order to obtain a coherent representation of the model's result we have obtained the importance of each feature and then we have trained multiple regression trees on a subset of the most significant features. The best performing trees have then been plotted and compared. An example of the results is shown in Figure 3, which portrays the path followed by the model in order to predict the habitat richness of a sample area in the test set.

The figure portrays the feature space of each node variable chosen by the tree, and how the data was split to obtain the binary division. This tree was trained used only the 10 most important features of the Bulgaria dataset, the maximum depth was set to 4 branches and its overall regression R^2 score on test set is equal to 0.76. Of course, a single tree is less accurate than the correspondent random forest model, but it is possible to inspect the tree by visualizing it. In this case the prediction for this test area was of 0.61 habitat richness. Three main variables

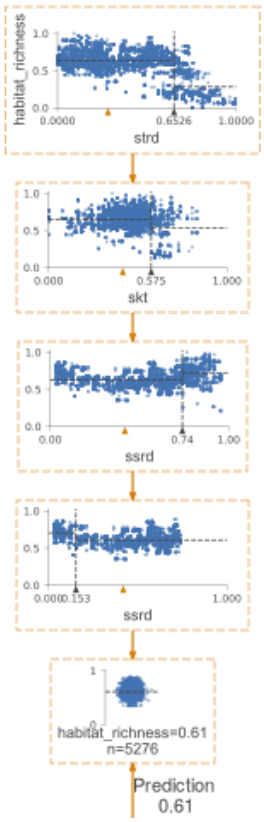


Fig. 3: Path followed by the Bulgaria Decision tree to regress a test sample

were considered by the model for the regression: strd (Surface thermal radiation downwards), skt (Skin Temperature) and ssrd (Surface Solar Radiation Downwards).

D. Images model performance

TABLE III: Score comparison of the images model over part of the Italy region

Method	R^2 ***Score***
Decision Tree	0.730
Random Forest	0.850
SVM	0.714
Neural Network	0.799
Images + Feature NN	0.784

As can be seen from table III the implementation of the images alongside the other feature has not be able to reach the basic Neural Network nor the Random Forest. This results can be justified by the small number of images that were not sufficient for the model to give better information than the ones already provided by the other features.

VI. Conclusion

Our experiment provides a reliable estimation of region-level species richness, using only features that can be sensed remotely by satellites and openly provided by the Copernicus program. We have evaluated the implication of area selection and how different features affect the results. We have also provided a way to explore and visualize the result of random forest by plotting the regression tree and the feature importances, which should be further integrated by ecological analysis.

Comparing our models results with similar experiments such as [11] and [12] we can say that the R^2 obtained in our experiment is generally higher, but some important considerations have to be made. For example, similar results to ours have been obtained on Mediterranean Climate in Chile by [12], in a $R^2 = 0.651$ was achieved using a dataset composed of LiDAR derived data and hyperspectral images, on a much smaller, habitat-level area, with higher spatial resolution than ours. We could compare this experiment with our results in the Italian area ($R^2 = 0.839$), whose habitats are fairly similar to the Chilean area considered by them in terms of altitude, climate, distance from the equator and distance from the see, however it should be stressed that to have a proper comparison we should run the experiment on the same area, with the same richness indicator. Another example could be [13], which achieved a $R^2 = 0.37$ in predicting the species richness of snails, bryophyte and vascular plants separately using a dataset composed of forest-specific and general purpose features with much higher resolution than ours. Validating our result by using the same richness index and on the same area of these papers should be the next step for our experiment, in order to understand the real effectiveness of our data choice with respect to theirs. Another important future step should be further investigating the effect of area size and internal composition of the dataset on the performance of the model.

Overall, this experiments proves that it is possible to obtain a reliable estimation of biological diversity entirely from satellite observations, and that further experiments on this topic could prove useful to obtain a coherent, scalable way of monitoring biodiversity on a regional scale. Moreover, we have provided multiple insight and visualizations of our results which could be useful for ecological analysis about the effect of each feature on the richness of species in forest habitats.

VII. Future Work

This work could be expanded in many directions. After further evaluating the models on different areas and different richness indicators, it could be interesting to experiment more with satellite images of different spectral bands, using the Neural Network architecture we have used in section V-D to produce even better performing models. Other interesting directions could be experimenting with different biodiversity indicators, for example regressing single species diversity or Shannon diversity index, or expanding the experiment to non-forest related species. Moreover, the dataset could be expanded including features from other Copernicus services in order to obtain a much needed monitoring of the status of biodiversity in marine ecosystems.

VIII. Contributions

Cosimo Chetta: Copernicus Climate Change data retrieval, Dataset cleaning, Regression models comparison, Grid Search, Image Network architecture

Vincenzo Madaghiele: Copernicus Land data retrieval, Biodiversity dataset research and retrieval, Areas research, Feature Importances, Tree visualizations, Image research, retrieval pre-processing and segmentation

References

[1] (2021) Regression of species richness biodiversity indicatorfrom satellite observations and environmental parameters, github repository. [Online]. Available: https://github.com/vincenzomadaghiele/Regression-of-biodiversity-indicators

[2] B. J. M. Anne E. Magurran, *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, 2011.

[3] B. M. B. et al., "The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models," *Methods in Ecology and Evolution*, vol. 4, pp. 327–335, 2013.

[4] G. BON. (2015) What are ebvs? [Online]. Available: https://geobon.org/ebvs/what-are-ebvs/

[5] W. D. K. et al., "Building essential biodiversity variables (ebvs) of species distribution and abundance at a global scale," *Biological Review*, August 2017.

[6] N. P. et al., "Satellite remote sensing, biodiversity research and conservation of the future," *Phil.Trans. R. Soc*, vol. 369, 2013.

[7] M. S. et al., "An applied statistical method to identify desertification indicators in northeastern iran," *Geoenvironmental Disasters*, 2018.

[8] "Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem," *Ecological Indicators*, vol. 60, pp. 870 – 878, 2016.

[9] E. O.-M. et al., "A comparison of artificial neural networks and random-forests to predict native fish species richnessin mediterranean rivers," *Knowledge and Management of Aquatic Ecosystems*, May 2013.

[10] A. K. et al., "Predictive mapping of reeffish species richness, diversity and biomass in zanzibarusing ikonos imagery and machine-learning techniques," *Remote Sensing of Environment*, Jan 2010.

[11] F. Z. et al., "Environmental predictors of species richness in forest landscapes: abiotic factors versus vegetation structure," *Journal of Biogeography*, vol. 43, p. 1080–1090, 2016.

[12] J. L. et al., "Comparing generalized linear models and random forest to modelvascular plant species richness using lidar data in a natural forest incentral chile," *Remote Sensing of Environment*, vol. 172, pp. 200–210, 2016.

[13] F. Z. et al., "Disentangling the effects of climate, topography, soil and vegetation onstand-scale species richness in temperate forests," *Forest Ecology and Management*, vol. 349, pp. 36–44, 2015.

[14] C. EU. (2014) Copernicus global land service. [Online]. Available: https://land.copernicus.eu/global/

[15] ——. (2014) Copernicus global climate change service. [Online]. Available: https://www.copernicus.eu/en/services/climate-change

[16] E. E. Agency. (2012) Richness of forest-related species and habitats indicator 2012 dataset, nov. 2018. [Online]. Available: https://sdi.eea.europa.eu/catalogue/srv/eng/catalog.search/metadata/81754d01-8bc3-49aa-a52c-86b3d212f94e

[17] C. B. Anderson, "Biodiversity monitoring, earth observations and the ecology of scale," *Ecology Letters*, vol. 21, p. 1572 – 1585, 2018.

[18] (2021) Regression of species richness biodiversity indicatorfrom satellite observations and environmental parameters, features analysis. [Online]. Available: https://github.com/vincenzomadaghiele/Regression-of-biodiversity-indicators/blob/master/DatasetCleaning/data_cleaning.ipynb

[19] E. K. M. et al., "Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories," *Ecology and Evolution*, Sept 2014.

[20] Scikit-Learn. sklearn.metrics.r2_score. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf