STEM Fellowship Big Data Challenge 2022

A Novel Method for Predicting Photovoltaic Potential in Canada

Abel Diress[1], Bilal Shaikh[1], and Ria Patel[1]

[1]Merivale High School

February 2, 2022

Abstract

To mitigate the effects of global climate change caused by fossil fuel emissions, Canada needs to reach net-zero emissions as soon as possible. But for a country that relies heavily on non-renewable resources to heat homes, fuel transportation and support industry, the renewable alternative must be reliable, efficient, and effective. One of the front-runners in sustainable energy solutions is solar power, and in a country as vast as Canada, it is sure to yield promising results. Our team analyzed the photovoltaic (PV) potential of different geographical sites across the country using data from the Canadian Weather Energy and Engineering Datasets (CWEEDS). Using K-means clustering, an unsupervised machine learning model, we placed all 564 locations into 5 clusters, and then predicted the PV potential for each cluster using a range of imminence and radiation variables. Through plotting our results on scatter graphs, we concluded revealed that PV potential in most of Canada is much higher than the world average (4.11-6.96 kWh/m^2). Furthermore, the province of Alberta - known for its tar sands and oil production - has the highest PV potential in the country, so the province has the potential to become the leader in solar energy production in Canada. These findings can aid provincial and municipal governments in optimizing their shift towards solar power and other renewable energy sources to maximize output. By identifying solar power as a strong alternative to fossil fuels, administrations can now start working towards setting up solar farms in places where they would optimally serve Canadians and take the first step to decrease our national carbon footprint.

Keywords
Canada, Photovoltaic Potential, K-means Clustering, Clean Energy

1 Introduction

In accordance with the UNESCO Sustainable Development Goal 7 "Affordable and Clean Energy", the profound global threat of climate change must be mitigated by a global shift towards clean energy. However, many question the viability and practicality of renewable energy sources over fossil fuels. That being said, climate change does not wait for public consensus. The time for change is now. This raises the question: what energy source is best suited to replace fossil fuels?

Of the various renewable energy sources on the market, solar energy is often the first choice, due to its current popularity (84% approval rate among Canadians) [1], and ease of access. However, despite its relative popularity, in 2019, only 2% of the world's energy output came from solar, compared to over 80% from fossil fuels [2]. In the face of rapid climate change, these metrics are lacking. The International Renewable Energy Agency estimates that by 2050, in order to maintain sustainable carbon dioxide levels, over 85% of the total energy produced must be renewable [3]. Considering solar power's current popularity, and the goal set by the International Renewable Energy Agency, solar energy production must reach a higher output.

Thus, with the goal of optimizing Canada's solar energy output in mind, our team explored the viability of solar energy in 564 individual regions across Canada. Viability, or photovoltaic (PV, units in kWh/m^2) potential, was calculated with regards to six attributes: global horizontal radiation, direct normal radiation, diffuse horizontal radiation, global horizontal illuminance, direct normal illuminance, and diffuse horizontal illuminance. Using these attributes, a K-means Clustering model was created to place the locations into groups of similar PV potential.

K-means Clustering is an unsupervised form of machine learning capable of grouping training data into a predetermined amount of groups using several variables. Our usage of a clustering-based algorithm, rather than a strict formula, allows our team's model to easily and effectively determine PV potential even in regions outside of the predetermined 564 Canadian areas with comparatively limited data, making this method globally applicable.

We hope that our study can be used as a guide for the Government of Canada in deciding where to facilitate and endow solar energy farms in the future. This study may also prove useful to average Canadian citizens, who, in an attempt to lower their carbon footprint, wish to install and optimize their own solar panels.

Figure 1: CWEEDS locations (2020 release). Locations with crosses are new the 72 new northern locations with this release.

2 Materials & Methods

2.1 Raw Data

The data used for this study was collected from the Canadian Weather and Engineering Datasets (CWEEDS), which included data from the Government of Canada, Natural Resources Canada, SUNY (the State University of New York), and NASA (the U.S. National Aeronautical and Space Administration). [4] [5] [6] The data was broken down into different locations around Canada, with greater coverage in more densely populated areas. The updated 2020 version of the CWEEDS used in this study contains data from 564 Canadian locations in all 10 provinces and 3 territories (see Figure 1). The time period for the data spans 29 years, from 1998 to 2017.

In machine learning algorithms, more data ensures a higher degree of accuracy in the model and a variety of data ensures the model isn't over-fitted. This dataset was chosen as it allows for a greater variety and more accurate coverage of the measurement of environmental factors in which PV potential is dependent.

2.2 Data Analysis

To determine the PV potential of each of the locations in the CWEEDS dataset, an unsupervised categorization algorithm (K-means Clustering) was used to categorize each of the variables into clusters representing PV potential. The chosen algorithm to determine the centroid in each 'learning' cycle for each cluster utilized the Euclidean distance of each point, expressed as:

$$dist(l_1, l_2) = \sqrt{\sum_{i=1}^{n}(v_i - u_i)^2} \qquad (1)$$

Where $(v_i - u_i)^2$ represents the square difference of each variable between two given locations.

From there, a variety of locations with predetermined PV potential were included in the clustering. Based on which cluster the predetermined locations were placed in, the PV potential range of all locations within the cluster were determined. To perform this, a variety of data visualization tools were used. A combination of the SciKit Learn machine learning library, Python, and Pandas was used to determine the clusters and their locations. For presenting these results, a geopolitical heatmap of Canada with the PV potential represented by tones of the colour was formed using Tableau and Microsoft Excel.

2.3 Variable Selection

Of the 45 variables in the CWEEDS dataset, 6 of the most reliable and available variables across the data set were selected for this study. Of the six, three of them are representative of the light received by the area (illuminance), while the other three are representative of the radiation received by the area:

- **Global Horizontal Radiation (GHR)**

  The total solar radiation incident on a horizontal surface. This metric is an important factor in determining PV potential for flat plate solar panels, the most common variety of panels for rooftops.

- **Diffuse Horizontal Radiation (DHR)**

  The solar radiation received by a panel that was diffused via atmospheric particles. Essentially, the excess radiation that lies unrepresented within the DNR.

- **Direct Normal Radiation(DNR)**

  The amount of solar radiation received by a solar panel, normal to direct sun rays. This variable, combined with DHR make up the global horizontal radiation.

- **Direct Normal Imminence (DNI)**

  The amount of solar illumination received by any given area, normal to direct sun rays.

- **Diffuse Horizontal Imminence (DHI)**

  The indirect solar illumination onto a specified area. This variable, combined with DNI, make up the global horizontal illumination.

- **Global Horizontal Imminence (GHI)**

  The total illumination received by a solar panel. This variable is a summation of DNI and DHI.

3 Results

3.1 Pre-Clustering

	GHR	DNR	DHR	GHI	DNI	DHI
count	564	564	564	564	564	564
mean	268.35	288.77	123.37	49.75	78.49	34.42
std	33.15	49.61	14.51	41.96	42.10	42.80
min	141.00	130.00	77.00	21.00	24.00	10.00
25%	247.75	261.00	115.00	31.75	55.00	15.00
50%	277.50	290.00	125.00	36.00	72.00	20.00
75%	293.00	323.00	134.00	50.00	91.00	33.00
max	321.00	392.00	158.00	579.00	583.00	573.00

Figure 2: Descriptive statistics for each of the six variables measured. The statistic 'std' represents one standard deviation of the array of variables. This data represents each of the non-zero variables for all 564 locations.

As seen in Figure 2, the Diffuse Horizontal Imminence has the highest standard deviation relative to its average (>100%). Therefore, we predicted that this value will play the most significant role in determining each location's cluster.

A further investigation via a geopolitical map (see Figure 3) presents the correlation of the DHI and the location.

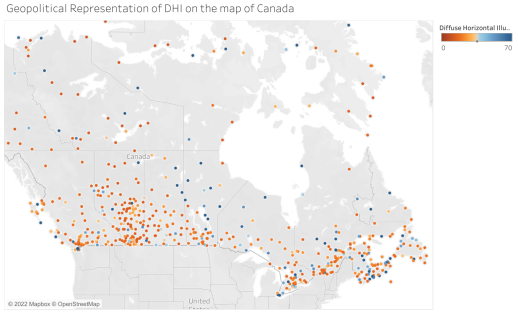

Figure 3: A representation of the Diffuse Horizontal Imminence in each of the 564 locations with their positions on the Canadian point mapped

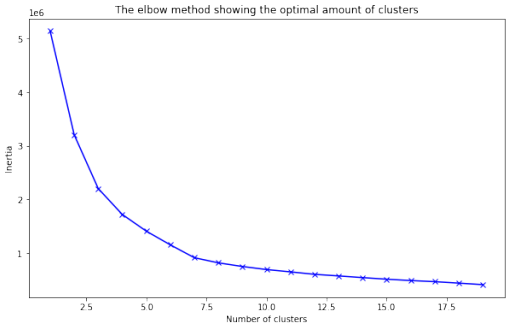

Figure 4: Correlating the number of clusters in a K-means model and its loss amount. This graph was formed using Matplotlib, Pandas, and Sklearn to calculate the inertia.

Figure 4 was generated to determine the optimal amount of clusters in the model. The graph represents 20 K-means clustering models with the number of clusters measured against their inertia. The inertia represents the Euclidean distance of each location in a specific cluster relative to the centroid of that cluster.

The inertia acts as a gauge of the approximate loss of the model based on the number of clusters. The goal of all models is to minimize the loss of the clusters. From this, it can be observed that the inaccuracy decreases significantly until 5 clusters, in which the difference of loss is minimal. Therefore, it can be concluded that 5 clusters will be the optimal amount of clusters for the model.

3.2 Clustering

After the K-Means model was created, all 564 locations were placed into 5 clusters. These clusters were numbered 0 to 4. However, at this point in the investigation, the PV potential range of each cluster was yet to be determined.

The difference between the size of each of the clusters is indicative of the uneven distribution

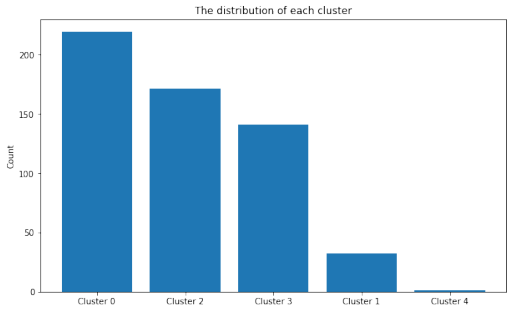

Figure 5: Bar graph representing the number of locations in each cluster.

of the locations, with a greater bias towards the southern locations. Therefore, remote areas, such as Collins Bay (Cluster 4), are more likely to be less accurate in their PV potential than major cities. The number of locations in each cluster follows a logarithmic trend, however, the data is not statistically significant ($p > 0.05$).

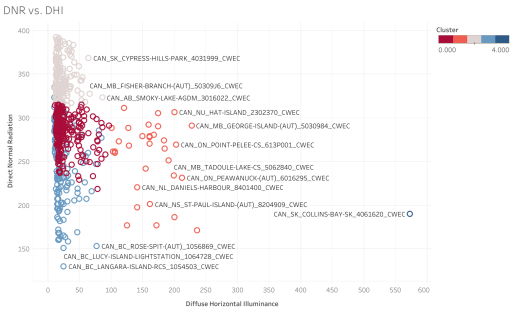

Figure 6: Comparing the Direct Normal Radiation (DNI) and Direct Horizontal Illuminance (DHI) with the locations coloured based on their respective clusters.

As shown in Figure 5, locations in the same cluster tend to have similar DNI and DHI. Noticeably, locations in Cluster 2 tend to have high DNR and low DHI. As a result, it can be assumed that this cluster will generally have a higher PV potential than the others. A high DNR and low DHI demonstrates that these locations tend to receive a high amount of direct solar radiation without diffusing it. Regardless of these observations, these two variables alone were not enough to predict PV potential, as shown in the fact that many of the clustered points overlap each other, leaving room for uncertainty between clusters.

Similarly to Figure 6, the scatter plot in Figure 7 shows that these two variables alone were not able to predict the PV potential of these Canadian locations. Once again, however, Cluster 2 tends to have the greatest ratio between solar radiation gained to solar radiation diffused.

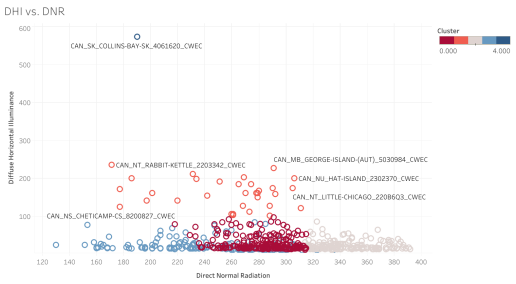

Figure 7: Comparing the Diffuse Horizontal Illuminance (DHI) and Direct Normal Radiation (DNR) with the locations coloured based on their respective clusters.

3.3 Model Analysis

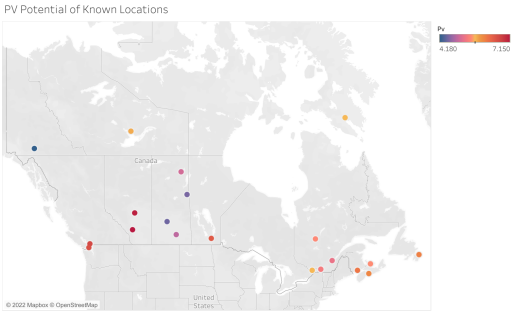

Figure 8: Mapping the locations with known PV potential, which will be used to determine the PV range of each cluster.

The geopolitical map in Figure 8 shows that the locations were dispersed across all clusters throughout Canada. Using this data, each cluster formed an average of the PV potential of all the locations with known PV in that cluster.

To find the range, we took half of the maximum and minimum known PV in each cluster resulting in the following PV ranges for each cluster:

- Cluster 0: 5.24-5.82 kWh/m^2
- Cluster 1: 4.11-4.55 kWh/m^2
- Cluster 2: 6.24-6.96 kWh/m^2
- Cluster 3: 5.00-5.62 kWh/m^2
- Cluster 4: 4.65-5.13 kWh/m^2

With these ranges determined, a final geopolitical map of each of the locations PV ranges was formed.

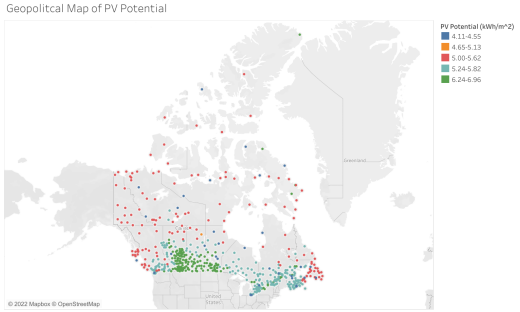

Figure 9: Mapping the 564 locations with their predicted PV potential.

4 Discussion

4.1 Concerns and Limitations

Our study contains two main areas of concern: the overlapping of clusters, and the lack of rural training data.

First, in our predictions, the PV potential range of Cluster 3 (5.00-5.62) overlaps with Cluster 4 (4.65-5.13). Although the overlap is minimal (only 0.13), there exists the possibility that some locations may have been put in Cluster 2, but their variables more closely resemble points in Cluster 4. However, such a case only applies to <10 of the total locations, making it a near negligible source of error.

Second, in training our model, urban data points were predominately used over rural ones due to the availability of data. As stated in our introduction, our study was meant to be used as a guide to the Canadian government in establishing solar farms across the country. We concluded that the areas in Canada with the highest PV potential are the prairie provinces. Generally, these areas were not close to large urban communities. It would be most logical for the Canadian government to build solar farms in sparsely populated flatlands. However, our model had mainly used urban data points as training data. The average urban setting varies vastly from its rural counterpart. Most notably, the DHI will differ, as in an urban setting there are many more sources that will reflect and diffuse direct solar energy. This variable is one which our model biases, thus, the calculated PV potential in rural areas may be inaccurate. Nevertheless, this flaw doesn't affect allows us to accomplish our tertiary goal with a greater amount of accuracy. Since the majority of the Canadian population reside in urban areas, citizens who hope to establish their own solar panels will find our study to be highly accurate.

4.2 Photovoltaic Potential

As evident in Figure 6, Figure 7 and Figure 9, Cluster 2 had the highest PV potential, with an estimated value of 6.24 to 6.96 kWh/m^2. The locations in Cluster 2 were localized in southern Alberta, Saskatchewan, and Manitoba. Past papers indicate similar results, with the Prairies being the most promising for PV potential [7]. This is a result of the high amount of solar radiation the area receives and the optimal temperature throughout the summer and spring seasons (20°C to 25°C)[8]. The cluster with the second-highest PV potential, Cluster 0, is localized in southern Ontario, southern Quebec, and southeastern British Columbia. Similar to areas in Cluster 2, these predominately southern locations are exposed to a significant amount of solar radiation.

The high PV potential in Alberta presents a possibility for a prosperous solar power industry in the area. Currently, Alberta's solar PV systems produce 736 megawatts (MW) of energy; compared with natural gas, which produces over 10,000 MW. [9] In addition, solar power represented only <0.1% of Alberta's total energy production. [10] Our team estimated that to produce the same amount of energy created by Alberta's natural gas, only 4,138-4,615 square meters of solar farms would be required.

This study's predictions diverge from past papers when modelling for Indigenous communities. Conventional prediction models indicate that there exists little to no PV potential in many northern Indigenous communities; however, our model finds there to be a promising amount. A possible explanation for this deviation could be that past papers tend to use solar-rooftop data [7], whereas the CWEEDS dataset used in this study covered most tillable land in each of the location using ground placed sensors. This means that our study was not restricted to areas with high levels of development and infrastructure to gauge PV potential. Nonetheless, this subcategory of the study could become a promising subsequent topic in the future.

Overall, the PV potential of most areas in Canada is much greater than the global average ($\tilde{4}.8$ kWh/m^2). This opens up many doors of opportunity for renewable energies in Canada, and helping combat climate change.

4.3 Other Prediction Models

As previously established, our team decided to calculate PV potential using K-means clustering, rather than a strict formula. Our results differed, particularly in northern Canada, where our model predicted greater potential compared

to other models. This is likely due to distribution of Diffuse Horizontal Illuminance (DHI). DHI is mainly determined by three factors, higher occurrences of clouds, a higher water vapour concentration, and a higher amount of atmospheric pollution [11]. These three factors are all plentiful in northern Canada. Since our model biases the DHI value, it will predict a greater potential in those areas than traditional methods. That being said, our model does have considerable advantages over other methods of calculating PV potential. One such advantage being the ability to include other variables, such as the angle at which the solar panel is tilted at, or the maximum power generated by a specific type of panel. In the coming years as consumers begin to adopt rooftop solar panels technology, these variables will be of great importance. [12]

Similarly, by using a clustering algorithm the variety of usable variables increases exponentially, whereas the strict formula is limited to a few predetermined variables. This is especially useful in areas with limited data as our model can be used as a rough estimate to determine if further investigation and data collection is warranted.

4.4 Future Work

A potential area of further investigation is the use of solar power in northern Indigenous communities. As shown in Figure 9, some remote areas in the Northwest Territories, such as the Iqaluit community of Pond Inlet, show high PV potential. A more thorough investigation of these Indigenous communities can prove to be useful source of industry. Currently, many of these areas rely on government subsidies for energy; however, solar power would provide a route of autonomy for Indigenous communities.

One component to further add to the prediction of PV potential of Canada is the economic investment into renewable energies, and solar energy in particular, in the studied locations. This would prove to be a more accurate gauge of what a given location can realistically implement in terms of solar PV systems as installation and management of solar panels is expensive.

In addition, our study implicitly assumes that solar energy and energy storage technology will maintain the same quality as when the data was collected (2020). Although this assumption is satisfactory for short-term prediction, it must be revised to predict solar energy production implementation in Canada in the long term (20-50 years). In order to do this, the K-means model would have to account for the rate in which several statistics in energy storage and production technology are changing, including by not limited to energy density, number of solar cells placed in an area, and efficiency of commercial solar panels.

Conclusions

Our study found that in areas with high total irradiance (GHI and GHR), the PV potential was higher. In Canada, the areas with the highest PV potential were the southern prairies, notably, Alberta. As is evident in the study's findings, Canada's potential is relatively high compared to the global average, especially for one of the northernmost nations. In the K-means clustering algorithm, an optimal 5 clusters were used, which had a calculated PV potentetial range of 4.11 to 6.96 kWh/m^2. Despite Canada's potential, in 2017, oil and gas contributed $128 billion to Canada's GDP. The non-renewable energy sector is essential to Canada's economy; however, by investing in renewable energy sources, particularly solar energy, Canada can achieve net-zero emission goals and become a front-runner in the renewable energy industry. Our study serves as a starting point, as PV potential will continue to increase in the coming years due to technological innovations increasing the viability of solar energy and other renewables. Canada is a huge country with ample room to support hydro dams, wind farms, and solar farms. By understanding where PV potential is highest, governments can be assured that their investments will pay off, and citizens can start relying on solar energy as a constant source of electricity, consequently helping the world fight climate change.

Acknowledgements

We would like to thank everyone at STEM Fellowship who supported us in this amazing journey of data science, including Ben Fedoruk, Anish Verma, David Babalola, Zackary Masri, and Vivek Saahil. We would also like to thank Valerie Han, Environment Canada, and the Natural Sciences and Research Council of Canada for providing us with the meteorological data necessary to conduct this study.

References

[1] Energy: Canadians tilt towards prioritizing renewables; one-third would split investments between green and oil. https://angusreid.org/energy-priorities-2021/, July 16 2021.

[2] Hannah Ritchie and Max Roser. Energy. *Our World in Data*, 2020. https://ourworldindata.org/energy.

[3] Adnan Amin. Renewables are the key to a climate-safe world. https://www.irena.org/newsroom/articles/2018/Nov/Renewables-are-the-key-to-a-climate-safe-world, November 16 2012.

[4] Environment Canada and Climate Change. Canadian weather energy and engineering datasets (cweeds), Nov 2021.

[5] Engineering climate datasets.

[6] Chun Yin Siu and Zaiyi Liao. Method for converting CWEEDs weather files to EPW format for multiyear simulation of building thermal dynamics. *MethodsX*, 7:101016, 2020.

[7] Fariborz Mansouri Kouhestani, James Byrne, Dan Johnson, Locke Spencer, Paul Hazendonk, and Bryson Brown. Evaluating solar energy technical and economic potential on rooftops in an urban setting: the city of lethbridge, canada. *International Journal of Energy and Environmental Engineering*, 10(1):13–32, November 2018.

[8] Environment Canada. Alberta - weather conditions and forecast by locations, Nov 2021.

[9] Green alberta energy: Statistics.

[10] Canada Energy Regulator Government of Canada. Canada energy regulator / régie de l'énergie du canada, Mar 2021.

[11] Global Solar Atlas. *What's the difference between DNI, DIF and GHI?*, 2017.

[12] Xiaoyang Song, Yaohuan Huang, Chuanpeng Zhao, Yuxin Liu, Yanguo Lu, Yongguo Chang, and Jie Yang. An approach for estimating solar photovoltaic potential based on rooftop retrieval from remote sensing images. *Energies*, 11(11), 2018.