Math INDUSTRY

2021 Final Report

Pacific Institute for the Mathematical Sciences
25 SINCE 1996

Math INDUSTRY

PREFACE

The second annual *Math*^{Industry} (Math to Power Industry) virtual workshop took place during August 2021, offered by the Pacific Institute for the Mathematical Sciences (PIMS), with the help of its training, industry and government partners. The workshop was launched in 2020 by PIMS in response to the economic impacts of the COVID-19 pandemic on mathematical sciences graduates. With the pandemic impacts continuing in 2021, and an ever-increasing need for mathematics and data sciences-skilled talent in North American industry sectors, online workshops like this one have become a vital bridge for connecting mathematics graduates and postdoctoral fellows to job opportunities in industry.

As in the first year, the workshop program began with a 10-day training bootcamp followed by a 2-week experience working as part of a team on a real problem provided by an industry or government agency partner. Workshop training courses included training on the latest programming and data workflow environments, effective teamwork, EDI (equity, diversity, and inclusion), communication, ethics in data science, startups and entrepreneurship. Seven of the real-world problems were contributed by industry partners, with four companies returning from the previous year, and three were contributed by municipal and federal government agencies. Motivation for these problems ranged from companies trying to make use artificial intelligence in their decision-making, improve the safety and operation of their products, minimize environmental impacts of food production and transport, manage and control insect infestations, and improve the health of our communities. Our teams approached these problems through advanced mathematical modelling, statistics, optimization, and computational techniques. In some cases the results had immediate positive impacts for these organizations, saving them both time and money, and in other cases the industry partner or government agency was able to use the workshop to recruit the highly skilled talent they need to be successful.

This *Math*^{Industry} 2021 Final Report compiles the results obtained by each of the ten workshop teams on their industry challenge.

ACKNOWLEDGEMENTS

We acknowledge the contributions and participation of everyone that helped make this event possible. A heartfelt thank you to:

WORKSHOP FOUNDERS

Dr. James Colliander, former Director of PIMS

Dr. Kristine Bauer, PIMS University of Calgary Site Director and PIMS Jobs Committee Chair

ORGANIZING COMMITTEE

Dr. Allen Herman, PIMS University of Regina Site Director
Dr. Matthew Greenberg, University of Calgary
Dr. Jolen Gallagher, University of Manitoba - Industrial Relations
Dr. Linglong Kong, University of Alberta
Ruth Situma, PIMS Program and Communications Manager

TRAINING PARTNERS

MITACS
CYBERA
WESTGRID
QUANTSIGHT

INDUSTRY AND GOVERNMENT AGENCY PARTNERS

AERIUM ANALYTICS LTD.
ATCO LTD.
CSTS HEALTH CARE LTD.
CITY OF WINNIPEG – INSECT CONTROL BRANCH
IOTO INTERNATIONAL LTD.
MCMILLAN-MCGEE LTD.
NATURAL RESOURCES CANADA
SERIOUS LABS, INC.
THEORYMESH INC.

TRAINING COURSE INSTRUCTORS/GUEST SPEAKERS

Dr. Thomas O'Neill, University of Calgary, ITPMetrics
Ian Allison, PIMS
Marie-Helene Buhrle, WestGrid
Dr. Ron Baecker, University of Toronto
Dr. Dhavide Aruliah, OpenTeams, Quansight
Laura Gutierrez Funderburk, Cybera
Lorena Solis, University of Calgary
Samantha Jones, University of Calgary
Dr. Amelia Taylor, Zymergen, Inc.

ACADEMIC MENTORS

Dr. Aysa Fakheri Tabrizi, PDF, University of Calgary
Dr. Joshua Brinkerhoff, University of British Columbia
Dr. Peijun Sang, University of Waterloo
Dr. Jonathan Gallagher, PDF, Dalhousie University
Dr. Slim Ibrahim, University of Victoria
Dr. Hui Huang, PDF, University of Calgary
Dr. Cüneyt Akçora, University of Manitoba
Dr. Shaun Lui, University of Manitoba
Dr. Andrii Arman, PDF, University of Manitoba
Dr. Adam Kashlak, University of Alberta
Dr. Julien Arino, University of Manitoba

INDUSTRY AND GOVERNMENT MENTORS

Heather Vooys, Aerium Analytics
Vakhtang Putkaradze, ATCO
Nisha Mohan, ATCO
Gianoulla Lakka, CSTS Healthcare
Ali Hashemi, CSTS Healthcare
Ken Nawolsky, City of Winnipeg
Jennifer Bodnarchuk, City of Winnipeg
William Spat, IOTO Intl.
Edwin Reid, McMillan-McGee
Devin Goodsman, NR Canada
Adam Preuss, Serious Labs
Chris Bunio, TheoryMesh

PARTICIPANT LIST

Andrii Arman
Joel Benesh
Noah Bolohan
Edgar Pacheco Castan
Carson Chambers
Adeyemi Isaiah Fagbade
Jonathan Gallagher
Sajad Fathi Hafshejani
Hamid Hamidi
Parham Hamidi
Rachel Han
Mitch Haslehurst
Jules Hoepner
Saimon Yeal Islam
Santanil Jana
Amit Jha
Aniket Joshi
Avleen Kaur
Bryan Kettle
Xiaowei Li
Guojun Ma
James McCurdy
Alexandra McSween
Pedro Jose Sobrevilla Moreno
Bahar Mousazadeh
Maksym Neyra-Nesterenko
Arnaud Ngopnang
Bo Pan
Thomas Pender
Igor Pinheiro
Mishty Ray
Eric Rozon
Sebastian Moraga Scheuermann
Natalia Accomazzo Scotti
Moumita Shau
Mahsa Nasrollahi Shirazi
Sam Simon
Ellie Thieu
Shen-Ning Tung
Yiyu Yang
Aidin Zaherparandaz

Contents

Aerium Analytics

AI for road surface scanning

Amit Jha, Avleen Kaur, Aysa Fakheri Tabrizi, Hamid Hamidi, Parham Hamidi, Xiaowei Li

Industry mentor: Heather Vooys

Abstract

This analysis is to develop a machine learning tool that detects cracks on road pavements by processing RGB aerial images of the road. The data considered is unlabelled. Thus supervised and unsupervised learning approaches are applied to design an intelligent detector. The results are combined to produce optimal feedback.

1. Introduction

Artificial intelligence (AI) has been a valuable mechanism for the growth of many industries, such as education, health care, lifestyle, transportation, web search engines, etc. It can be applied to enforce road safety by monitoring the condition of the roads, which is also the goal of this project. Since roads are the primary mode of transportation, they are essential for everyone. The materials mainly used for the construction of roads are concrete and asphalt. The road pavements can develop defects such as cracks and potholes due to fatigue from excessive usage, extreme changes in weather temperatures, water accumulation, etc. Such defects are treated according to their severity to ensure safe travel. Otherwise, they could degrade the quality of roads and cause more irreversible damage. Traditionally, the inspection of pavements is done by a team physically. Or a human expert analyzes the aerial images of road pavements for surveillance.Since there could be numerous pictures for a particular province, having a human expert go over each one of the images is a laborious task. It could also pose a financial strain on the government.

This project develops a machine learning or computer vision tool, which reads the images of road pavements. It highlights the defects that are present on the pavements in the output images. The data provided for training and testing the machine learning model consists of RGB aerial images of road pavements of a parking lot and other roads. Such a model is highly beneficial for the government, as it reduces the need for human efforts. Consequently, the model is economical and is a much faster option. Time and money are two key factors that affect the performance of a task in the real world. On reducing these stressors, the government gains more resources to enforce measures for the correction of road pavements efficiently, see [1], [2].

The data provided for designing this model is unlabelled. It leads to two possibilities. The first one is labelling the data manually for designing this project, which is a supervised learning approach. Otherwise, we could train the model by using unlabelled data itself. It is called an unsupervised learning approach. We explore both of these approaches in this project to build a coherent tool. However, there are only fourteen images in the data set. We generate more images for increasing the data by using the techniques such as image segmentation and windowing, which split each image into several parts. The first stage creates several sub-images for the input image, and then each one of them is processed by the model. The resultant output images of each sub-image are combined back together to give the input image back in its original form at the end of the processing. Since the images are aerial, they contain several other elements such as cars, markings on the roads, poles, trees, shadows, etc. They all contribute as noise in data, which is the biggest challenge in developing this model. Moreover, the defects can be of a random shape, so it is difficult to train the model to detect such defects in the pavements. Some similar problems on detecting cracks have been studied in [3], [4], [5], [6].

This report is organized as follows. Firstly, section 2 presents a supervised learning model that highlights cracks on aerial images of road pavements is presented. The second approach, an unsupervised learning technique is depicted in section 4. The combination of the two techniques is described in section 5, Finally, section 6 discusses the results and conclusions of this study.

2. Approach 1: Supervised Learning

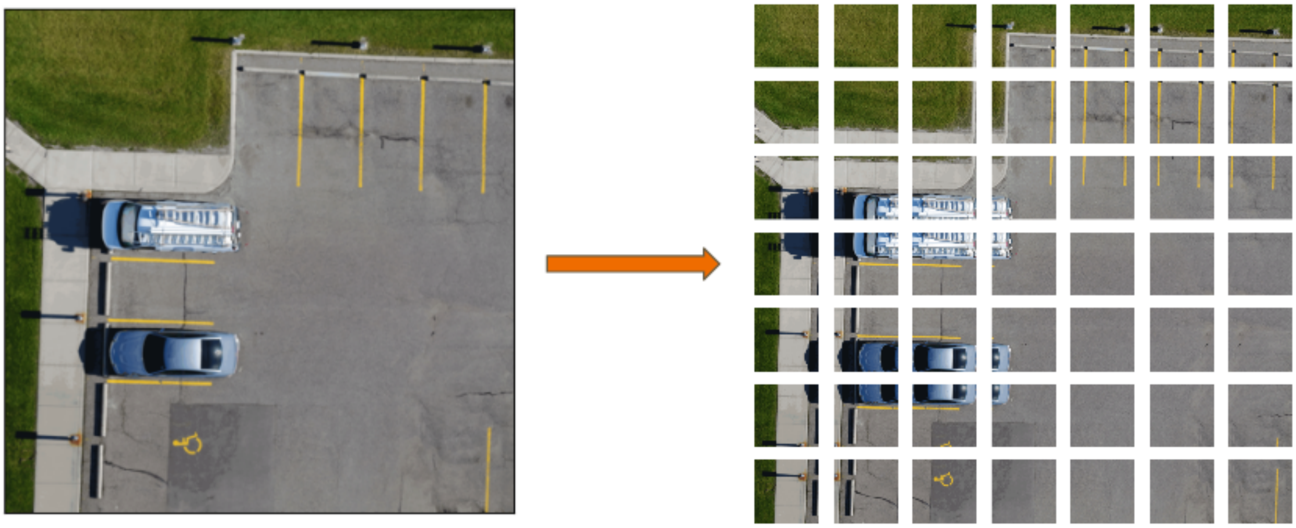
Background

Convolutional Neural Network (CNN) is known for its exceptional preformance on image recongnition. Our idea here is to train a CNN to "identify" the cracks in pictures. However, typical data sets for training CNNs usually contains tens of thousands of samples with labels. Here we have only 14 images and no "label" for the cracks. Moreover, a few images should be left out of training to use as test set for the trained model. In order to increase the sample size and label them, we do some pre-processing on original data set as detailed below.

Methodology

Creating tiles and overlapping patches

The original dimension of the pictures were mostly 3724 by 3724 pixels (with a few pictures having even more pixels). We turned each picture into overlapping tiles/patches with dimension 224 by 224 pixels. Overlapping tiles helped us capture each crack better (in case the boundry of a tile falls on a crack). Moreover, it helped us increase the data for validation and training.



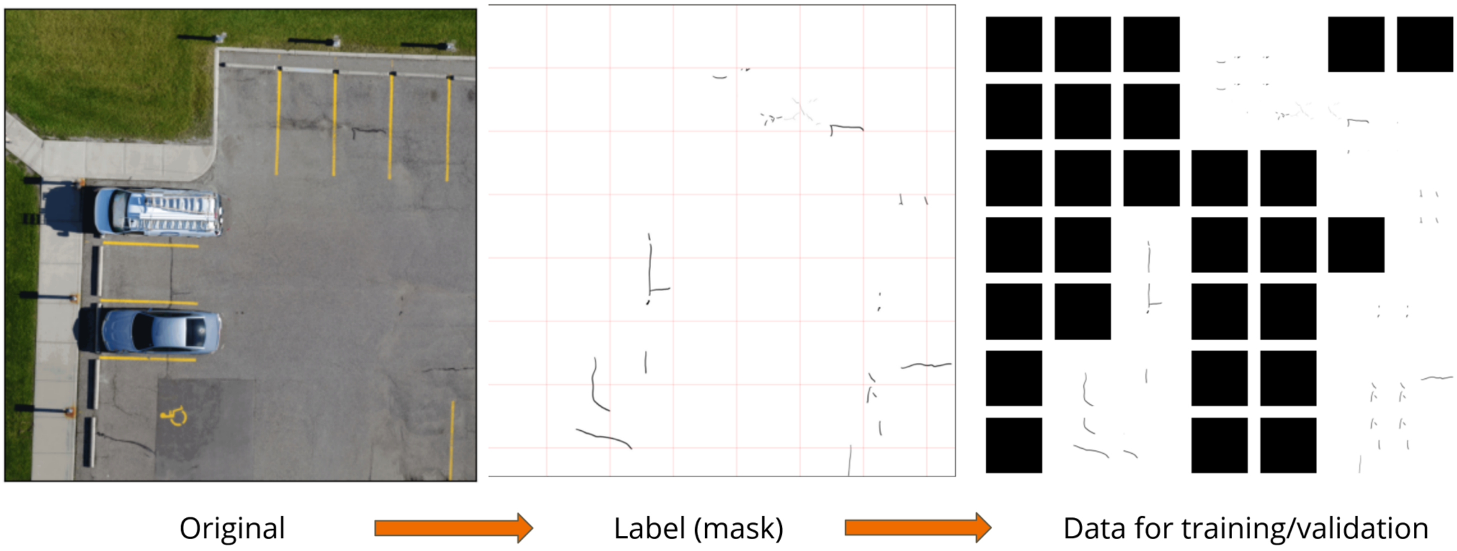
Labelling

The data that was provided to us by Aerium Analytics was not labelled. This is often the case in the industry. To train a Neural network with supervised learning, we needed to label our raw data. Our first task was to identify the cracks which we want to use for training our Neural network.

Before we start labelling, we had to decide what are the cracks we wanted our Neural network to detect. To do this, we looked at some of the previous worked that has been done, as well as a pavement distress identification manual [7]. Some of the cracks in our data were miniature and hence we decided that they are irrelevant for our training of crack detection.

After we came to consensus about the cracks we wanted to consider, we used a free graphics editor, called GIMP, to mask over the cracks. This was done by drawing over the sensible cracks with a digital pen. Then, the masks were turned into an appropriate file format (PNG) to be read by OpenCV.

We also formulate the problem into a binary classification. Using the masks we draw, each tile was labelled as positive or negative depending on whether the tile contains more than 800 pixels of cracks (`PIXEL_THRESHOLD=800`).



We note here that the actual data for traning and validation is different than the above figures. The above figures are only for demonestation of the process.

Training and validation set

We took 11 out of 14 images and created 4169 patches with about 15% overlap between adjacent tiles, among them 842 were labelled positive (i.e., containing cracks). Note that this is an unbalanced data set (about 20% samples are positive) and we will later address this issue in the loss function for training. We further randomly divided these 4169 patches into training set and validation set with a ratio of 6:4. During training, the training set is also augmented with a random horizontal and/or random vertical flip to obtain up to 4 times training samples.

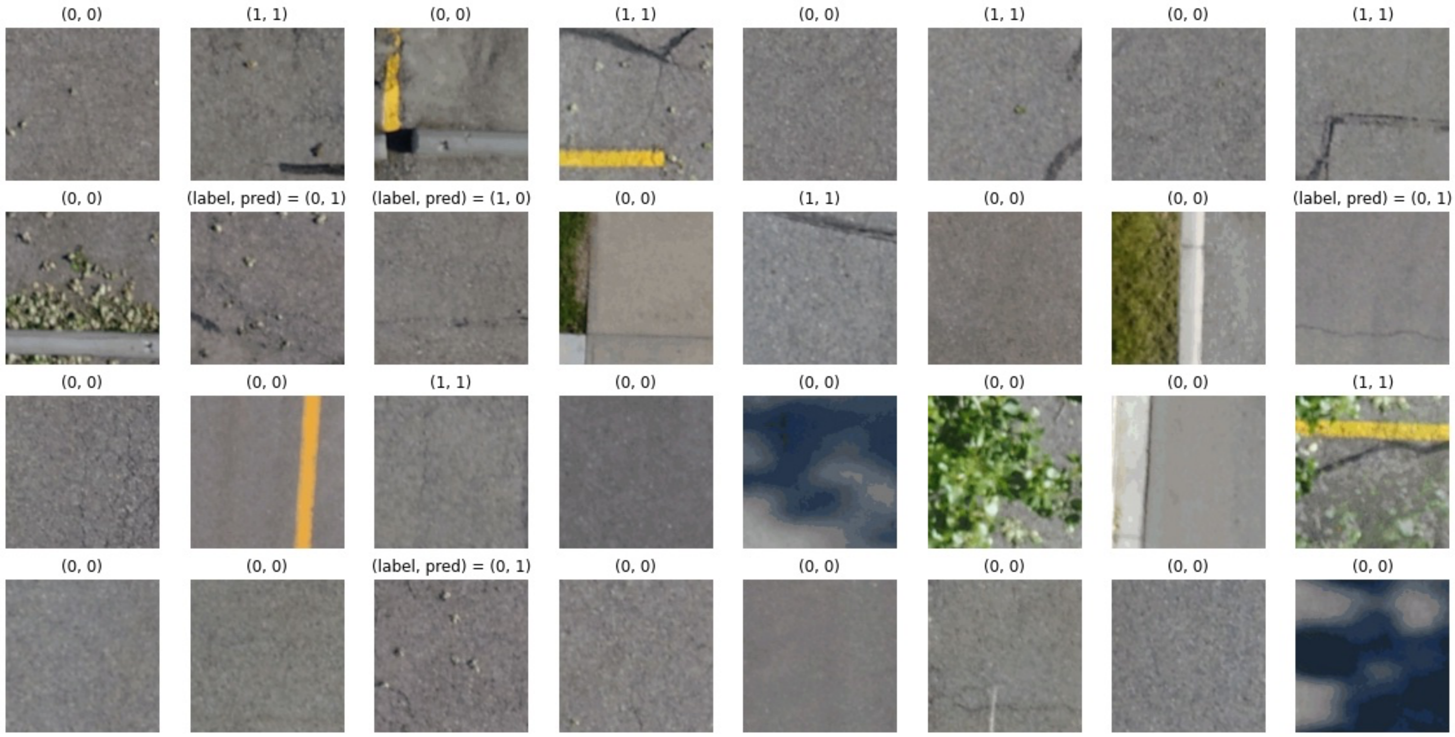
Test set

For the remaining 3 images, we turned them into patches of size 224 by 224 pixels with (almost) zero overlap. They will be used to show our results.

Training CNNs with transfer learning

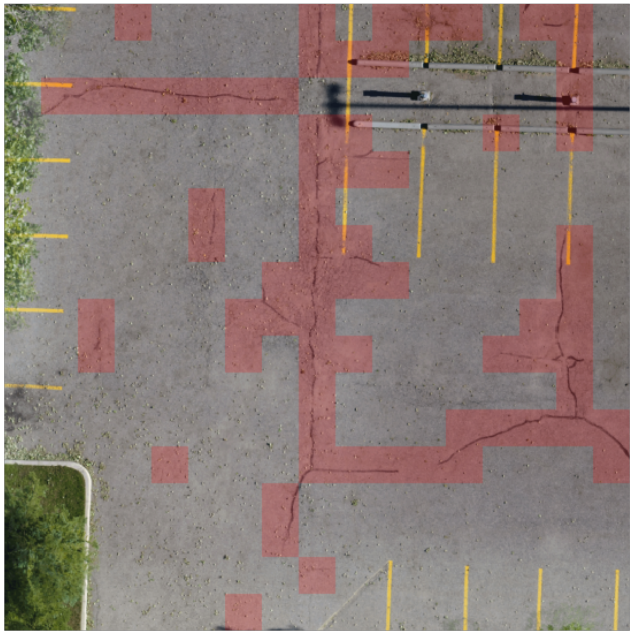
To further combat the issue with small data set, we use pre-trained networks as the backbone of our classifier. In particular, we took ResNet50 [8] and modified its last layer (with an added `Dropout=0.4`) to perform binary classification. The training process started with pre-trained weights (except last layer) and 'fine-tuned' the entire network. The loss function is the standard binary cross entropy, but with 4 times the weight on positive samples to 'balance' the data set. The optimizer is stochastic gradient descent with momentum. We ran this for up to 20 epochs and evaluate the accuracy on validation set. At last, we obtained a classifier with about 94% accuracy on the validation set, and the number of false positive and false negative samples on validation are 50 and 39 respectively. The training is done using Google Colab and the details can be found in this Colab notebook.

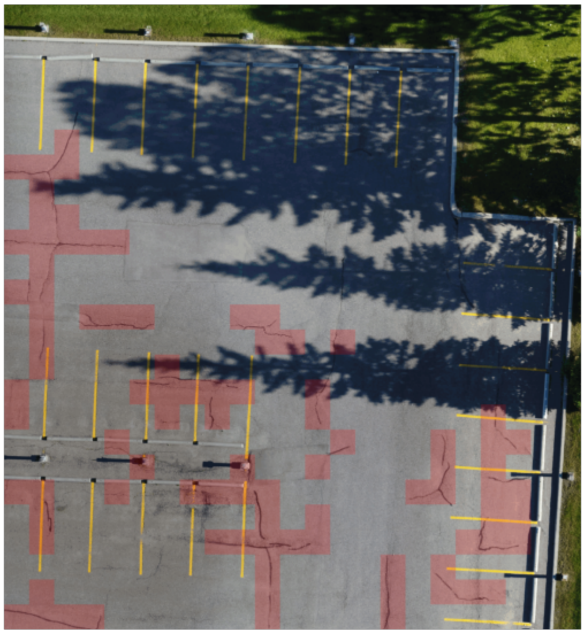
Below are some examples for this classifier's prediction on validation set.



Results

Below we show the predictions of our classifier on the 3 test images. Here red shaded patches are the ones our classifier thinks there is a crack in them. For the first image, there are no cracks in the image and no cracks are predicted. For the second image, the classifier is able to identify almost all patches with cracks, aside from a few false positives. For the third image, the classifier again identified most of the cracks. However, it missed almost all of the cracks in the shadow of trees. We believe this phenomenon is mostly due to there being very little such samples in traning set, and the model can be further improved in this scenario with more related training samples.





Further Remarks

- Overfitting
  During training, the ResNet50 model was able to reach an accuracy of above 98% on the training set while having about 94% accuracy on validation set. We also tried the same training approach on ResNet18 and ResNet101, and obtained similar accuracy on training and validation set. However, as we can see from the prediction examples on the validation set (e.g., last picture in the second row), our labelling itself is not 100% accurate. Thus we suspect there is some overfitting in the model(s) at this point.
- Freeze all but last layer
  We also tried freezing all the pre-trained network parameters and only re-train the last layer. In this case, the trained model performs a few percentage worse on the validation set when compared to the 'fine-tunning' model above. We believe one reason for this could be that there is not enough 'free' parameters to capture the complexity of training set. One future direction may be only to freeze some of the bottom layers and 'fine-tune' the upper layers. This could also help with the overfitting issue since now there are fewer parameters to train.
- Different patches sizes
  We also tried larger patch sizes (e.g., 448 by 448 pixels, 600 by 600 pixels, etc.) and found similar accuracy on training and validation set. In the end, we choose the smaller patch size because it can give more accurate location about where the cracks are.

3. Approach 2: Unsupervised Learning and Image Processing

Background

As the number of samples is limited, we considered unsupervised approaches simultaneous to the supervised approach. We applied unsupervised techniques consisting of two independent sections. One is the classical image processing techniques used in finding edges in images, including but not limited to filtering, morphological operations, and edge detection. However, the cracks with the edges of other objects such as cars are detected in our image processing steps. To overcome this problem, we tried another approach whic is clustering pixels using machine learning techniques. We used two different clustering algorithms to identify objects and remove unwanted parts of the image. In addition to object detection, clustering could be used to group pavements' pixels as a unique cluster among others. In the following, the classical and machine learning approaches would be described in detail.

Image processing techniques

The images contain pixel information as the values in different channels such as RGB, standing for Red, Green, and Blue, respectively. The pixel values in each channel could take a value between 0 to 255, which would be the intensity of that specific channel. To identify a crack in an image based on the channels' intensities, we have to suppress other features or highlight the feature of interest, the cracks, to an extent where it is easily identifiable by the edge detection method.

Image processing techniques mainly consists of 4 steps to delineate crack patterns:

1. Size reduction
2. Noise reduction and filtering
3. Edge detection
4. Morphological transformation

Size reduction

In our dataset, the images have the size of order 3000×3000. To reduce memory consumption and perform our operations faster, we reduced the sizes of the images to 800×800

Edge detection

The points (pixels) that have discontinuities in brightness are detected by mathematical operations as the edges in the image. Here, we applied the Canny edge detector to find the patterns of pavements in our dataset.

Morphological transformation

In the edge detection process, some points can be missed and cracks would be shown as discontinuous patterns. Therefore, morphological operations such as closing and erosion have been used in the last step of crack detection using image processing techniques.

Shadow removal

This is a complex problem that depends on many factors and camera settings. We tried implementing the paper by [9] which talks about shadow removal. The shadows can be removed if the log response of camera sensors is projected along an illumination invariant vector. The grayscale results were not very promising. We tried to improve the results provided at Stack..

Machine learning approache

We used two different algorithms for clustering the pixels, K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

In the following, steps of both these appraoches and their results will be explained.

Methodology and Results

Noise reduction and filtering

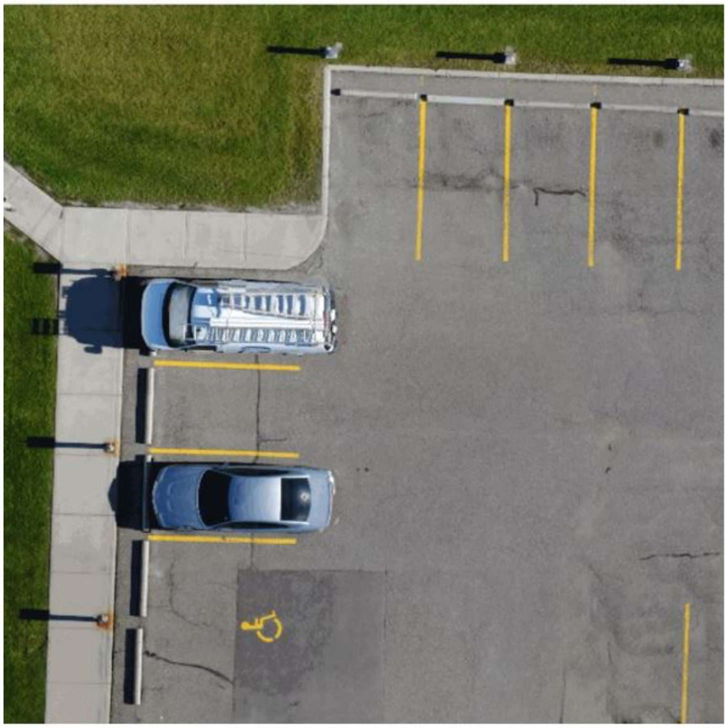
This is the first step in removing unwanted information from the image. Filters are mainly used to smooth images without losing edge precision. Here, we used distinctive filters that cut the high frequency by convoluting a kernel over the pixels. We used bilateral, median, Gaussian, and uniform blurring filters on images. Bilateral and gaussian blur preserve edges if appropriately applied.

We tried following sequences of operations on raw images.

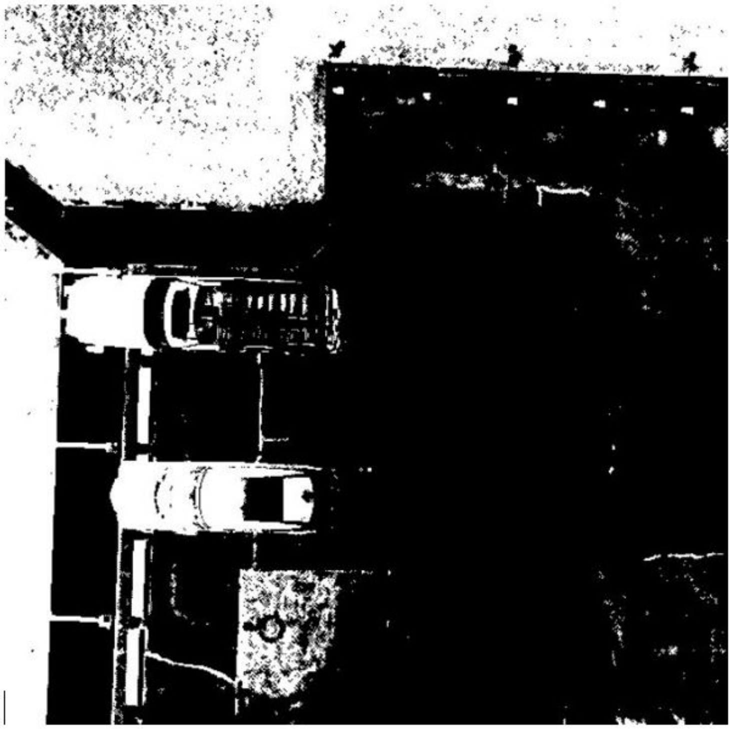
Method 1

We used bilateral blur on gray image followed by convolution with various kernels. After that we applied optimal threshold image binarization Otsu's algorithm. The results are the following:

- *RGB >Gray > Bilateral blur filter > Sharpen convolution > OTSU thrsholding*.

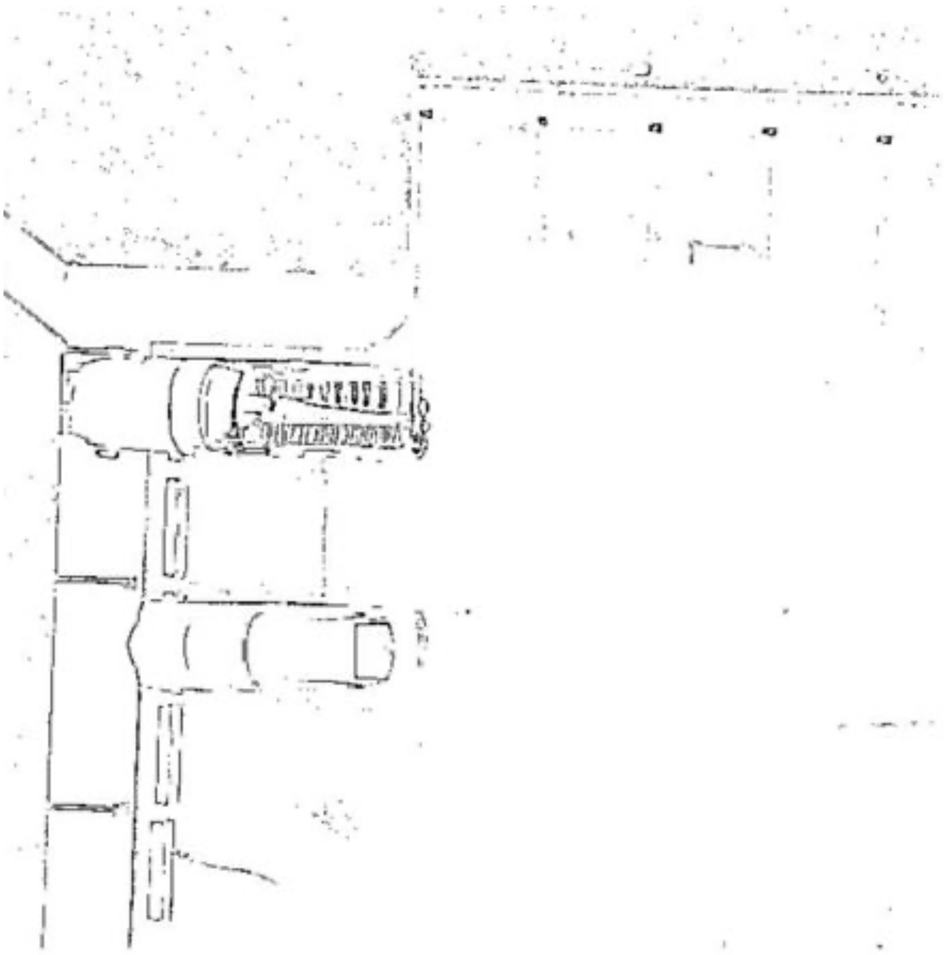


Original image



Bilateral blur + Sharpen + Thresholding

- *RGB >Gray > Bilateral blur filter > Laplacian convolution > OTSU thrsholding*.



Median blur + Laplacian + Thresholding

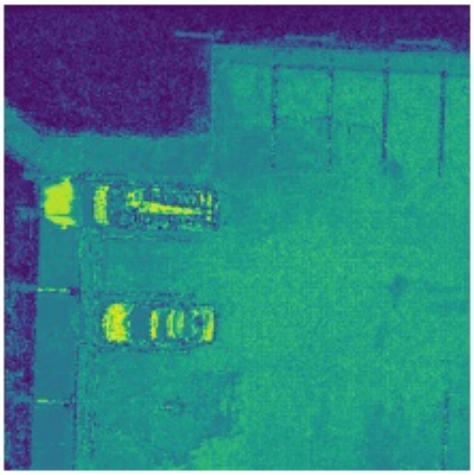
Method 2

We did some and operation value and saturation layers followed by thresholding.

- *RGB > HSV > Bitwise_and(V-Constant,S) > Otsu's thresholding*.

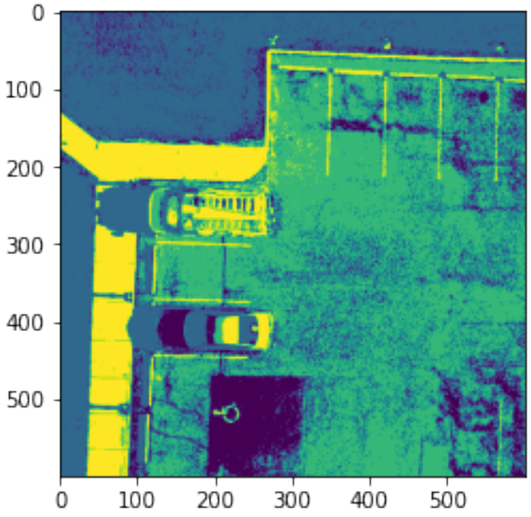
Method 3

We performed DBSCAN on the resized images without applying any filters. As shown in the following figure, the cracks can have a distinctive cluster among other clusters.



Method 4

We performed K-means on the resized images without applying any filters. As shown in the following figure, the cracks can have a distinctive cluster among other clusters. Also, other objects such as cars and grass can have a distinctive cluster.



4. Combined Approach

Based on the nature of supervised and unsupervised learning methods and their application in our problem and based on our observations from the experimental results, each of our two proposed methods has its own advantages and disadvantages in solving the crack detection problem.

Approach 1 advantages:

- Getting most out of the existing data by labeling them and using the labels in training the model.
- Capability of eliminating the redundant objects in the picture such as trees, cars, and curb sides and better performance in distingushing between the shadows and cracks.

Approach 2 advantages:

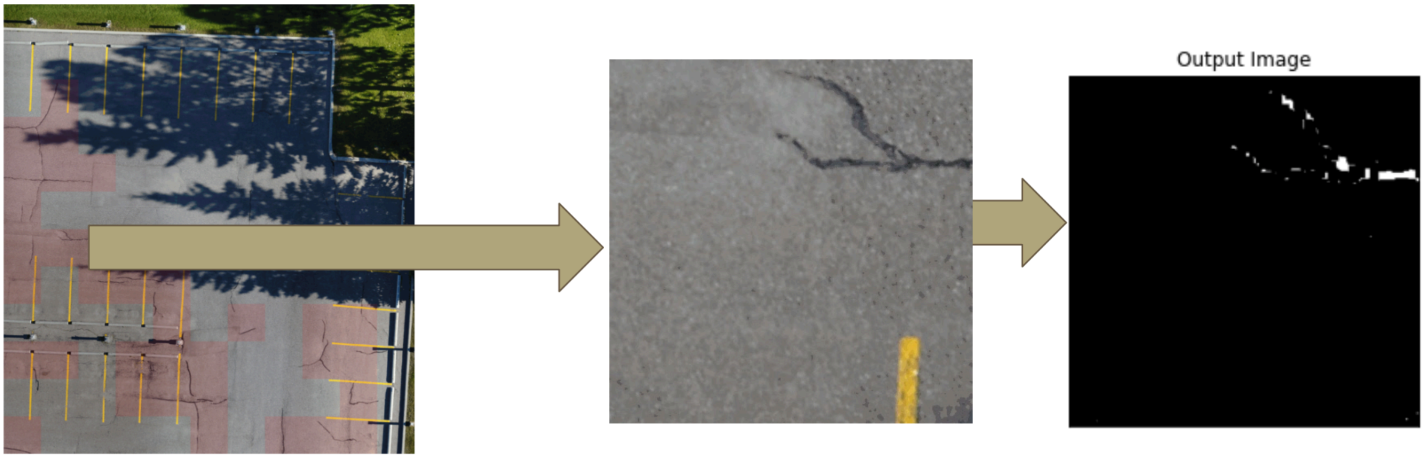
- Ability to detect the exact shape of the cracks vs finding the small patches that contains cracks.

Methodology

In order to benefit from the advantages of both models, we combined our two methodologies and created a framework that first, using our customized supervised learning model, detects small areas in the picture that contain cracks; then the detected area is given to our proposed unsupervised learning crack detector to output the exact location and shape of the crack. By this approach our framework is able to first eliminate the areas that doesn't contain any cracks and output the exact shape of the cracks.

Results

The following figure shows the result of applying our combined method to one of the case studies. The left picture is the output of oursupervised model in which we have detected the patches that contain cracks and highlighted them in red. Each of these red patches are given to our unsupervised model. In the picture one these patches that is predicted as contining crack is selected and is shown in the middle. The supervised model has correctly detected this patch as a patch with crack, however it doesn't output the exact crack shape. The right picture in this figure shows the output of feeding this patch to our unsupervised model. It can be seen that the unsupervised model further refined the output and and eliminated the yellow line and provided the exact shape and location of the crack.



5. Conclusion

In this project, we have developed a machine learning tool that detects cracks on road pavements. We have used unlabbeled data provided by Ariem Analytics and proposed a an intelligent detector that uses supervised and unsupervised learning aand computer vision methods to detect the cracks on the road pavements.

References

[1]. Government of Alberta, Annual Report Transportation 2020-2021. 2021. link

[2]. Transportation Alberta, Highway Maintenance Guidelines and Level of Service Manual, 2000. link

[3]. Yang C, Chen J, Li Z, Huang Y. Structural Crack Detection and Recognition Based on Deep Learning. Applied Sciences. Volume 11, Issue 6:2868, 2021. link

[4]. Suguru Yokoyama, Takashi Matsumoto, Development of an Automatic Detector of Cracks in Concrete Using Machine Learning. Procedia Engineering, Volume 171, 2017, Pages 1250-1255. link

[5]. Arun Mohan, Sumathi Poobal, Crack detection using image processing: A critical review and analysis. Alexandria Engineering Journal, Volume 57, Issue 2, 2018, Pages 787-798. link

[6]. Arvydas Rimkus, Askoldas Podviezko, Viktor Gribniak, Processing Digital Images for Crack Localization in Reinforced Concrete Members. Procedia Engineering, Volume 122, 2015, Pages 239-243. link

[7]. Minnesota, Department of Transportation, Office of Materials and Road Research. Pavement Distress Identification Manual. link

[8]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[9]. Finlayson G.D., S. D. Hordley, C. Lu and M. S. Drew, On the removal of shadows from images. The University of East Anglia, Norwich, UK., and, Simon Fraser University, Vancouver, Canada.

[10]. T. Ahmed Mahgoub Ahmed, Zhangcan Huang, Fan Xi, Liu Hai Ming, Detection crack in image using Otsu method and multiple filtering in image processing techniques, Optik – Int. J. Light Electron Opt. 127 (3) (2016) 1030–1033.

● ✕

Math INDUSTRY

FEASIBILITY OF USING VERTICAL AXIS WIND TURBINES (VAWT) PLACED ON HIGHWAY MEDIANS

MENTORS: NISHA MOHAN, JOSHUA BRINKERHOFF
ALEXANDRA MCSWEEN, JAMES MCCURDY, ADEYEMI FAGBADE, MAHSA N. SHIRAZI

Abstract. As concerns grow over limited resources, fluctuating energy prices, global warming, and environmental degradation, the need for renewable energy sources has becomes paramount in maintaining and meeting the current energy demands. Among the renewable and clean energy technologies, wind energy is one of the most efficient and cost-effective sources of renewable energy production, costing 1-2 cents per kilowatt-hour after the production tax credit by governments. While natural wind speeds over various continents in the world span from 0 to 20 m/s, Vertical Axis Wind Turbines (VAWT) placed on highway medians make it possible to utilize consistently higher wind speeds due to vehicle motion. Additionally, the energy generated by these wind turbines is reported to increase multi-fold due to the shearing winds generated on both sides of the medians by the on-going traffic, and offer a great opportunities to reduce costs as well as carbon footprint. The objectives of this project are to characterize vehicle-induced wind patterns to assess and empirically quantify the potential for wind energy-based electric generation for highway services, such as lighting and signage, in areas for which the electrical grid is either unavailable or for which interconnection would be complicated and expensive. In addition, we seek to optimize the number and positioning of VAWT turbines to achieve optimal results using criteria determined to be important, such as output power, ease of installation/repair, proximity to consumers, additive effects from positioning e.t.c. We seek to apply a computational fluid dynamics approach, real-life traffic, geographical and weather data and subsequently investigate the economic feasibility of implementation of this technology.

Contents

1. Introduction to the Project

The utilization of wind energy is technologically driven and has gained significance as one of the best alternatives to the traditional methods of generating electricity with a viable potential to mitigate and reduce global warming effects on the environment. As a result, wind turbines have evolved technologically over the years with considerable potential to generate and contribute to the global energy productions as it represents a low-carbon alternative to conventional power-generation technologies that depend on fossil fuels. Wind energy is trending as one of the main sources of clean, renewable energy that would allow a rapid transition away from current fossil-fuel-based energy. However, if wind power is to drive a significant share of global energy supplies, the efficiency and energy density of wind turbines need to be improved. Despite the urge potential, harnessing the power of wind turbines depends on many variables ranging from environmental factors to design specifications. One typical way of improving wind turbines power is through highway installation, particularly at a high wind gust zone. The deployment of wind turbines on the highways has a considerable potential to generate electricity from traffic-generated turbulence between the turbine and passing vehicles.

While travelling along the highway at high speeds, vehicles produce large amounts of wind energy in their wakes which is not being captured. By placing vertical axis wind turbines (VAWT) on highway medians, we could capture some of this energy and use it to power lights, charging stations, or other projects. In this project we identified factors that affect the performance of VAWTs, computed estimated power output on the 401 Highway and identified some other potential sites and future directions.

2. Turbine characterization and cost

The optimal performance of a wind turbine depends on the type, design, and structural specifications of the turbine among other factors. As a result, many different designs of wind turbines have emerged over the years. However, contemporary wind turbines can be identified based on the shaft orientation and rotational axis as either the horizontal-axis variety (HAWT) or the vertical-axis design such as the Savonius type, the Darrieus type, or the Giromill type(VAWT).

This project focus on the feasibility study of vertical axis wind turbine due to the marginal cost of the turbine and the ability to accept turbulent flow by air displacement from vehicular movement irrespective of the wind direction. VAWTs are designed to capture the wind kinetic energy irrespective of wind orientation, thus, setting aside the need for re-positioning along with the wind and thereby offering great benefits in places where the wind direction keeps varying. VAWTs are characterized by lower aerodynamics noise and fit in more readily into urban settings. This structural configuration considerably reduces the operating costs while improving stability and reliability. In addition, VAWTs have distinct operating features such as the ability to operate under irregular wind flow, slow cut-in speed, and low maintenance cost than comparably sized axial flow turbines. Based on the literature survey in support of this project, we identified Banki and Colite DS series turbine systems as potential wind turbines that can be placed on highway medians for power generation. Besides harvesting wind omni-directionally simultaneously, they are better suited for large array installation, that is, more able to stage multiple turbines on medians to capitalize on synergistic fluid interactions between the turbines. Collectively, these characteristic features imply that arrays of VAWTs can possibly gain power densities an order of magnitude higher than those of isolated wind turbines. However, harnessing wind power from any direction to create energy using these turbines is an attractive option and despite the intensive research in the field, the underlying performance parameters of the turbines and the question of scalability are not yet settled. One of the few studies that investigated the performance augmentations of pairs of VAWT was that of Dabiri [3], which were experiments in a desert with six $10m$ tall times $1.2m$ diameter VAWTs were conducted. The experiments investigated the effects of turbine spacing and direction of rotation. It was observed that while HAWTs experienced an overall decrease in power by $20\% - 50\%$ when placed in close proximity to each other (1.65 turbine diameter separation), the VAWTs enhanced the overall performance by $5 - 10\%$. In this regard, Banki and DS series wind turbines can be paired and interact synergistically to enhance the total power production when placed in close proximity.

The Banki and DS series turbines have cut-in wind speeds of about $1-3ms^{-1}$ and rated wind speeds that are less than $15ms^{-1}$. Table 1 provides more information on the specifications of the four models that were used as a basis of wind data interpolation for power generation.

Table 1. Colite VAWT specifications.

Model	**Banki**	**DS-300**	**DS-1500**	**DS-3000**
Cut-in wind speed (m/s)	1	< 3	< 3	< 3
Rated Wind speed (m/s)	10	13.5	12	12
Maximum Speed (m/s)	12	15.5	15	15
Rated Output (kW)	0.2-0.8	0.3	1.5	3
Rotor Diameter (m)	1	1.24	2.8	4
Rotor Height (m)	0.1	5.06	6.9	8.16
Turbine Cost ($)		3,585	18,825	26,625

3. The Data

In this project we identified several factors that could positively and negatively affect VAWT power output. For example, it is known that VAWTs perform better in laminar flow, and that the wake from vehicles is highly turbulent. Moreover, in [7] they showed that the power output from cars and SUVs is essentially negligible when compared to the power output from trucks. In light of this, we realized that since most trucks travel in the right (slow) lane, it made more sense to consider placing VAWTs on the shoulders of highways. Thus we only considered one way traffic. The East-bound direction on the 401 had 21% more traffic than the West-bound direction, and was used for this study. It should be noted that the wind generation modeling results presented in this report are based on a single VAWT. When considering VAWTs deployment, it should be recognized that the wind power technology is scaleable and thus multiple turbines can be deployed over a small area to produce additional electricity.

3.1. **The 401 Highway.** The 401 is the largest highway in Canada. Exploring data from [6], we found that in 2006, nearly 8000 vehicles travel on the 401 at Keele St every hour. On weekdays, nearly 2000 of those are trucks.

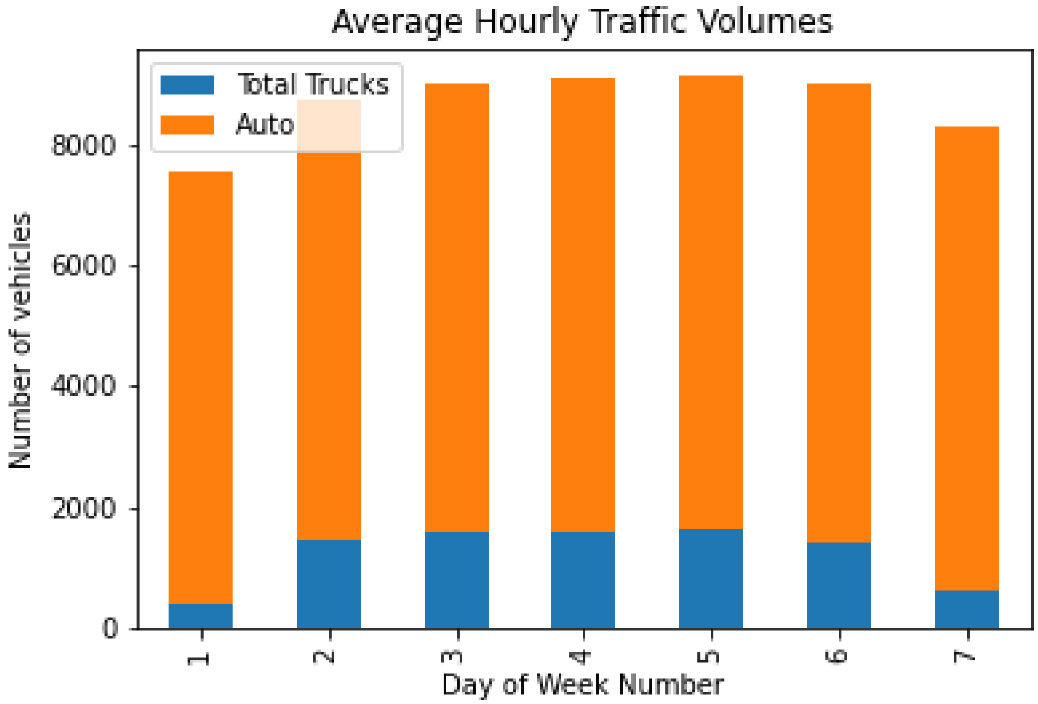


Figure 1. Average hourly traffic volumes on the 401 at Keele St made using Python.

To improve this data for our computations, we scaled it up by 12% to match population growth in the Greater Toronto Area. Next, we wanted to understand the lane by lane composition of the traffic. Based on [1], in a 6 lane highway about 50% of trucks are in the rightmost lane, while only 5% are in the leftmost lane. Extrapolating this for an 8 lane highway, we assumed about 40% of trucks are in the rightmost lane and 5% are in the leftmost lane. We applied this to our data set. Finally, the data from [6] was taken hourly over a two week period. We wanted to add in some seasonality to reflect the fact that there are more passenger vehicles on the road in the summer and less in the winter. We assumed the two week period took place in the Spring/Fall, repeated the data to have values for each day of the year, increased car volumes by 10% for days between June and September, and decreased car volumes by 10% between January and March, while keeping truck volumes constant.

This area of the 401 is a good candidate site because it has the largest truck volumes in all of Ontario. Moreover, due to the proximity to Lake Ontario there are higher windspeeds than in other areas. However, a metropolitan area has access to many other sources of power already, and the influence from buildings and other roadways may dampen the amount of wind the VAWTs can harvest. This analysis suggests that potential energy generation would vary from location to location and based on many factors, including the number of traffic lanes, median or center barrier type, traffic volume, vehicle type, and season.

3.2. **Other Potential Sites.** The data in [6] contains hourly traffic volumes for all main highways in Ontario. This would be a good data set to compute power output for other sites in Ontario. We also considered traffic data from the BC 5 highway which was reported on the Government of British Columbia website. We selected the daily data from 2018, and approximated some missing data by averaging the values from the previous years, we didn't have enough time to do the similar analysis we did for the 401 Highway, but it is a good candidate for future research on this project.

While other provinces did not have as robust traffic data, there are many other sites that could be good placement for VAWTs. For example, any areas with high wind speeds, in energy deserts, or with reasonable weather (not too much ice or snow) would make good candidates. However, in any case, we found the most important feature is high truck volume.

3.3. **Wind Data.** The National Renewable Energy Lab has developed the wind toolkit database [4] for onshore North American winds from 2007 to 2014. Wind speed, wind direction, and other meteorological are available at 5 minute intervals over a 2km x 2km grid. Each variable is available at various heights throughout the atmospheric boundary layer from 10m to 200m above the Earth's surface. Examining the wind data set, it is obvious that the turbulent wind flow exhibits high temporal and spatial variability. Therefore, the magnitude of a turbulent flow is best measured by maximizing data resolution (i.e., minimizing the sampling intervals). As a result, 10m wind velocity for the most recent year (2014) was used for this analysis. The 10m wind velocity was further scaled down to 2m height via a log-wind profile with a mean canopy height of 1m, this reduced the mean wind velocity by 40% from 2.95 m/s to 1.76 m/s. Furthermore, the wind was assumed to be constant over each 5 minute interval. Despite the wind and traffic data being from two different years, they were assumed to be independent, and no further adjustments were made.

4. Modelling

4.1. **Traffic Simulation.** The hourly truck and passenger vehicle rate on the 401 were each treated as a Poisson parameter and decomposed to the expected number of events per 2 second interval. If a truck, being a longer vehicle, was present within the 2 second interval, no other vehicle could be present; however the number of passenger vehicles, being smaller, were unrestricted. While there is is a non-zero probability that more than 1 truck is present within the 2 second interval, given that the average truck rate was 0.14 trucks per 2 second interval, there is less than 1% chance that the mean interval would contain more than 1 truck.

4.2. **Wind-derived Power.** As referenced in [7], the most applicable vertical axis turbines for are solonius or banki style turbines. The turbines studied are approximately 1m tall and 1m in diameter. While larger turbines certainly exist, sight-lines, safety, and space concerns limit the size of the turbine.

Given that the air is not moving at a standstill after passing through a turbine, a turbine doesn't capture 100% percent of the potential energy. Under ideal conditions a wind turbine captures between 30 and 50% of the potential wind energy. A typical manufacturer's rating for a 1m x 1m turbine is 300W at 10m/s, half of the 600W potential energy.

Unfortunately, in reading many small turbine reviews, the manufacturer's rating does not match up to reality. A variety of other conditions, notably turbulent wind can greatly decrease the power production. Engineering design studies [5, 8] refer to a 1m x 1m turbine with a manufacturers rating of 300W at wind speeds of 10m/s produces a maximum of 40W under turbulent wind conditions.

Furthermore, a feasibility study in Kuwait [2] measured the wind produced by vehicles along a highway to be 5.1 m/s. For a 40W turbine, with a 1m/s cut-in speed, this corresponds to 6W, which is the power production found in a truck found in [7]. Note that [2] suggests using a 3m high and 1.8m diameter turbine, they found this turbine to produce 50W at wind speeds of 5m/s. However, for this analysis, we will use the 1m x 1m turbine, producing a maximum of 40W at wind speeds of 10 m/s, the power curve used for this turbine is taken from [8, 5] and is shown in figure 2.

Research [7] finds that a single vehicle creates two gusts of wind, one from the air being pushed by the front of the vehicles, and another from the air being sucked in behind the vehicle as it passes the turbine. In this model, a single vector of wind, parallel to the flow of traffic was created over each 2 second interval. The wind vector was then added to the prevailing wind taken from [4]. The new wind vector was then converted to power using the power curve illustrated in Figure 2.

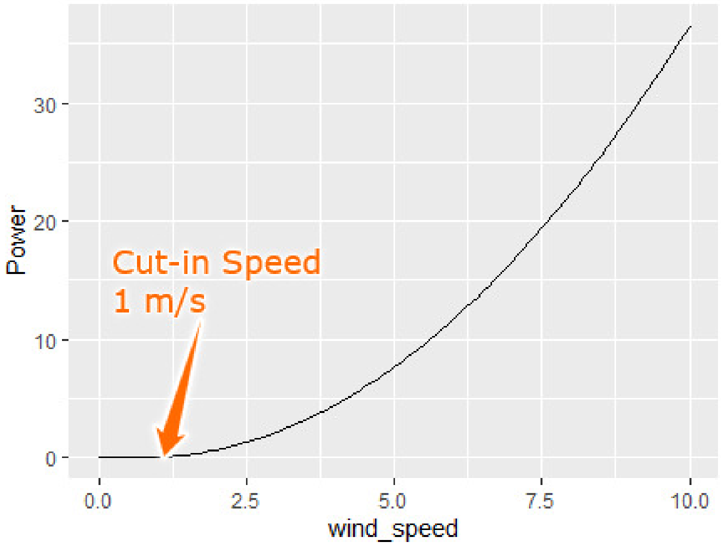


FIGURE 2. Power curve of 1m x 1m turbine in turbulent Conditions. The turbine has a manufacturer rating of 300W at 10 m/s, but experimental data suggests a peak of 40W.

	Single Vehicle	Two Vehicles	Three Vehicles	Theoretical Limit
Truck	6.48	9.71	11.32	12.92
Car	0.49	0.74	0.86	0.98

TABLE 2. Average Power(W) from a single turbine from multiple vehicles in series used in the vehicle-derived model.

4.3. **Vehicle-derived power.** Instead of modelling the wind velocity, we can model the energy produced. Research [7] shows that a truck can produce an average of 6W, with a peak up to 15W. Passenger vehicles produce a much smaller power output; however in the presence of other vehicles, they can contribute up to 1W of power.

To model this output, a variety of power decay curves were evaluated. Linear decay was found to overestimate the power, as multiple vehicles passed the turbine, it would produce higher and higher amounts of power. Exponential decay successfully reduced the maximum power output, but produced long tails under 1W of power. As a result, a combination exponential decay with rate 0.9 and linear decay with rate -1W/2s interval were chosen. The exponential reduced the peak power, while the linear decay reduced the length of the tail. Figure 3 illustrates an example of this power generation, and table 3 shows the mean power generation of multiple vehicles in series. In this model, the wind was treated independently to the power, and power generation due to the prevailing wind was added as per Figure 2.

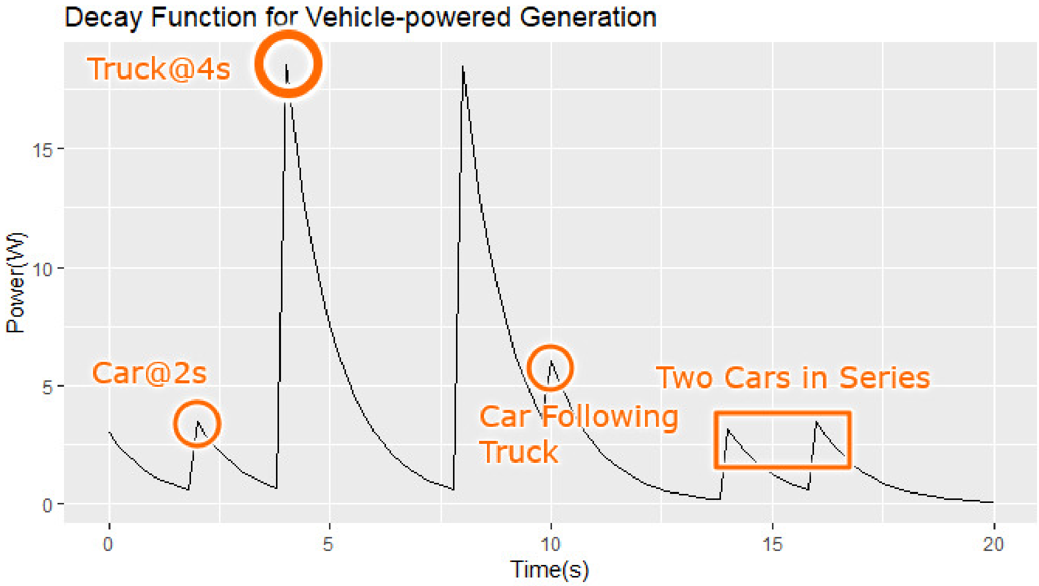


FIGURE 3. Vehicle Power generation used in the vehicle-derived model. As each vehicle passes the turbine, it produces a jump in power, which then decays according to a linear combination of exponential(0.9) and linear decay(-1W/2s interval).

5. RESULTS

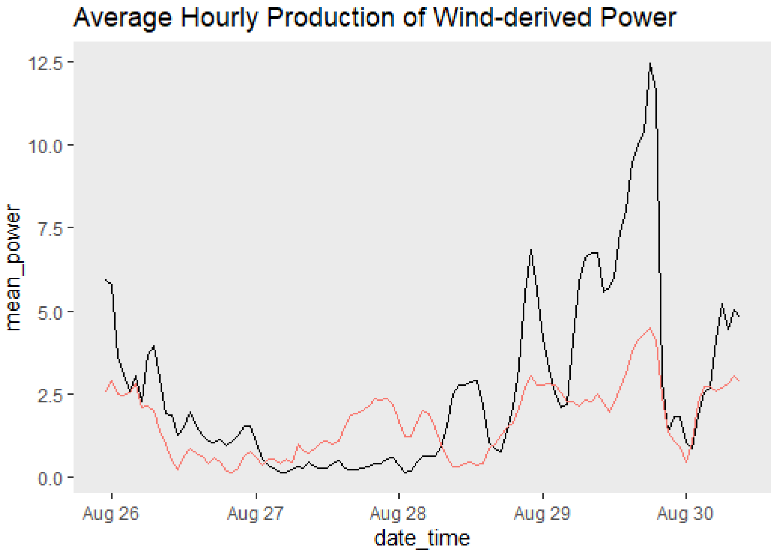


FIGURE 4. Sample of hourly wind-derived power production from August 26 to August 30. The mean prevailing wind speed (m/s) is overlaid in red

	5th Percentile(W)	Mean(W)	95th percentile(W)	Peak Hour(W)
Vehicle-derived Power	0.34	3.88	11.48	17.68
Wind-derived Power	0.05	5.98	21.75	34.83

Table 3. Mean annual power production from a single turbine from Vehicle and Wind derived power models.

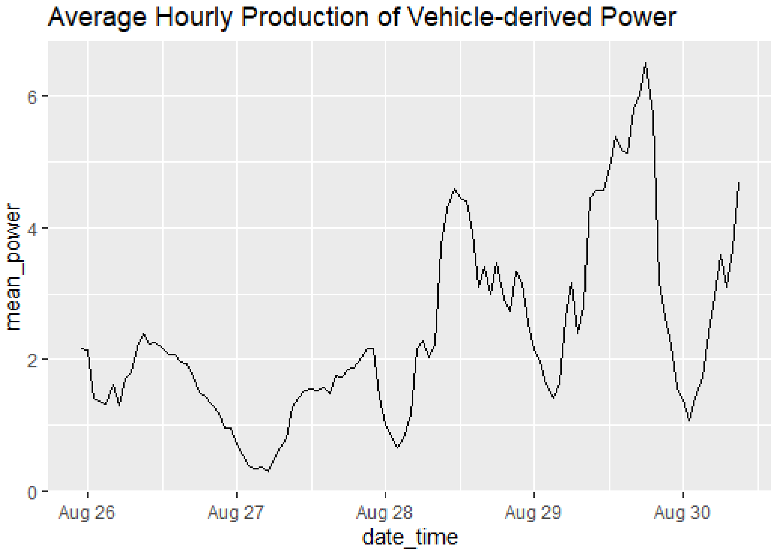


Figure 5. Sample of hourly vehicle-derived power production from August 26 to August 30

Table 3 shows the The wind-derived power production model produces on average 5.98W of power per year, while the vehicle-derived produced an average of 3.88W. The wind-derived model also had higher peak hourly power production. However, it had a lower 5th percentile, due to the model assumptions, where prevailing wind and vehicle produced wind were treated as independent vectors. If the prevailing wind was against the flow of traffic, the resulting vector would be low in magnitude, resulting in low power production. Figures 4 and 5 show a sample of 4 days from the year simulated. Increased traffic during the day creates increased traffic flow and thus higher power. Overnight fewer cars pass by the turbine, and the power production decreases accordingly. Figure 4 shows an overlay of mean wind speed in red, the higher windspeeds between August 29 and August 30 contribute to the increased power output.

6. Conclusions

Overall, we see a mean power production of 3.88W with the vehicle-derived power model, or a mean power production of 5.98W with a wind-derived power model. Unfortunately, this cannot power a light bulb, nor a Tesla. Given the size of the turbine studied, the potential for wind power generation is limited, as noted above, for a 1m × 1m block of wind moving at 10m/s there is a maximal potential energy of 600W. The most effective turbines can potentially capture half of that power, with turbulence and other conditions reducing that again. the CFD studies in [7] show that solitary truck produces 6W of power, and so a model producing an average of 4-6W of power, given the overnight lulls in traffic, could be seen as reason to study this further. If a larger turbine was selected, capable of producing more power at a lower speed, a series of turbines could produces a significant amount of power. However, turbine power scales linearly with area, so a much larger turbine would be necessary. Nevertheless, the generated data facilitates the correlation of energy potential to traffic flux, providing a basis for determining an optimum placement of wind energy generating systems along highways.

There is plenty of future work to continue from this project. This study neglects to take into account the effects of other lanes, the potential for trow way traffic, nor the effects of the surrounding building on wind flow. There were also many approximations during the process of turning the summary data into granular 2 second intervals. The traffic speed was assumed to be constant, thus neglecting the effects of rush hour and traffic jams.

Furthermore, the stochastic modelling of the traffic was relatively simple. Anecdotal evidence of driving on highways suggests that traffic comes in flows, with gaps between slower traffic. Modelling the traffic using Markov chains with a higher probability of passenger vehicles being stuck behind trucks could lead to a more accurate model. Additionally Figures 4 and 5, look like a stochastic process. Using these outputs and designing an optimal trading strategy for battery storage or selling back to the grid would be an interesting follow-up question. In addition, we could also investigate the optimum layouts of small vertical axis wind turbines and perform parametric optimization of three-dimensional analysis of dynamics fluid body interaction for a Banki rotor pair in various configurations. The performance augmentation of turbine pairs would help to estimate the possibility of boosting the derived wind power on highways.

Acknowledgement

We would like to thank Nisha Mohan from ATCO for creating this project and providing guidance throughout. We would also like to thank Dr Joshua Brinkerhoff from UBCO for providing resources and insight. Finally, we would like to thank PIMS for organizing the Math Industry workshop and allowing us to participate.

References

[1] Hamid R. Soleymani et al. *Truck Lane Distribution Factor for Alberta Multi-Lane Highways*. `https://ctep.ca/wp-content/uploads/2016/11/Truck-Lane-Distribution-Factor-for-Alberta-Multi-Lane-Highways.pdf`. 2006.

[2] Ehab Hussein Bani-Hani et al. "Feasibility of highway energy harvesting using a vertical axis wind turbine". In: *Energy Engineering* 115.2 (2018), pp. 61–74.

[3] J. O. Dabiri. “Potential order-of-magnitude enhancement of wind farm power density via counter-rotating vertical-axis wind turbine arrays”. In: *Renewable Sustain energy* 3.4 (2011), pp. 43–104.

[4] Caroline Draxl et al. “The Wind Integration National Dataset (WIND) Yoolkit”. In: *Applied Energy* 151 (2015), pp. 355–366.

[5] Ayman A Al-Maaitah. “The design of the Banki wind turbine and its testing in real wind conditions”. In: *Renewable energy* 3.6-7 (1993), pp. 781–786.

[6] MTA. *2006 Commercial Vehicle Survey: traffic volumes at survey stations.* `https://open.canada.ca/data/en/dataset/849fecf2-f166-499a-b473-5cfc7c7976ef`. Accessed: 2021-08-15. 2006.

[7] Saeed Nazari. “Power Generation from Localized Wind Energy on Highways using Vertical Axis Wind Turbines”. MA thesis. Okanagan, BC: University of British Columbia Okanagan, 2020.

[8] AR Winslow. *Urban Wind Generation: Comparing Horizontal and Vertical Axis Wind Turbines at Clark University in Worcester*. 2017.

E-mail address: `AM:amcsw087@uottawa.ca JM:michael.mccurdy@ucalgary.ca`

CSTS Healthcare

Natalia Accomazzo[1], Bo Pan[2], Eric Rozon[1], Ellie Thieu[3], and Yiyu Yang[2]

[1]University of British Columbia
[2]University of Alberta
[3]Amherst College

September 9, 2021

Abstract

Cancer treatments have been developed to give the 'precision medicine' paradigm of cancer therapy for years. While we have had some success with specific cancers, many other patients are put on a roller coaster of emotion through cycles of remission and relapse. At this point, there is abundant scientific evidence that tells us cancer is driven by multiple genes working in concert. Thus, due to the complexity and heterogeneity of tumor, the design of personalized therapies that are unique to every individual are suggested. To date, precision oncology trails have been performed and, unfortunately, several of these trials have been hindered by very low matching rates. The reason of it is commonly because of the use of limited gene panels, restrictive matching algorithm, lack of drug availability and clinician's knowledge. Based on our previous work, we have developed a computational system that can help to identify a personalized cancer therapy for every cancer patient, given their unique set of DNA and RNA. For each patient, our system identifies the best of set of target genes and active hallmarks, each of which have sets of available corresponding therapies associated with them. In a clinical setting, however, a clinician prescribes the therapy they believe is most appropriate for a given patient. In this paper, we would like to construct a set of similarity measures that allow us to compare our Aiomic therapies with those actually given by oncologists and we are interested in the evaluation of the association between the adoption rate of Aiomic therapies and clinical measurement, which provide us the evidence about the performance of the system and help to make the improvement accordingly.

1 Introduction and Background

Cancer is a disease that affects 14M people each year. While we have had some success with specific cancers, many other patients are put on a rollercoaster of emotion through cycles of remission and relapse. At this point, there is abundant scientific evidence that tells us cancer is driven by multiple genes working in concert.

Traditionally in medicine, drugs are designed and approved by testing on large populations. This type of results in only a portion of the population actually responding to a given therapy. However, different patients respond differently to the same drugs. Typically, on a large double-blind clinical trial with thousands of patients and two arms, everyone on one arm receives the same therapy A, while everyone else on the other arm receives exactly the same therapy B.

In the past several decades, we have learned that cancer is a highly heterogenous disease with multiple genes involved in cancer progression, and therefore requires combination therapies tailored to each individual's life history and genomic profile. Based on multiomic profiles one can model the cancer biology: which genes are driving the cancer, and which of the 10 hallmarks of cancer are active.

This allows the design of personalized therapies that are unique to every individual. However, this also presents a novel statistical problem, called the N-of-One problem, where it is difficult to achieve statistical power. Because each patient in a trial is receiving a differing, unique therapy, these therapies are not obviously comparable. One recent study, the I-PREDICT trial attempted to provide a comparison of personalized combination therapies when the therapy was only partially adopted. They did not directly evaluate personalized therapies, but simply determined that when a given therapy targeted more mutations, it was correlated with better outcomes.

2 Knowledge Gap and Problem Statement

We have developed a computational system that identifies a personalized cancer therapy for every cancer patient, given their unique set of DNA and RNA. For each patient, our system identifies the best of set of target genes and active hallmarks, each of which have sets of available therapies associated with them. In a clinical setting however, a clinician prescribes the therapy they believe is most appropriate for a given patient. In this context, if there are 100 patients, there are 100 unique therapies. Moreover, the therapy a patient receives may not match what our system identified as the best therapy. This problem then breaks down into the following sub-problems:

- We would like to construct a set of similarity measures that allow us to compare our Aiomic therapies with those actually given by oncologists.
- Given the entire set of patients, can we quantify adoption rates of Aiomic therapies
- Can we measure outcomes for Aiomic therapies as to whether they succeeded, partially succeeded or failed?

In contrast to the I-PREDICT approach, for our problem, we are interested in not just matching gene targets, but evaluating the success of Aiomic therapies when the given therapies are not an exact match. The problem can be stated as follows:

- How can we compare Aiomic-Therapies to Given-Therapies? For example, would it simply be the % intersection between the set of targets and hallmarks that are covered by the drugs in each therapy?
- Across all patients, given the set of Aiomic-Therapies and Given-Therapies, can we say how often Aiomic therapies were adopted? To what degree?
- Across all patients, given outcomes, when a Given-Therapy is not exactly the same as a Aiomic- Therapy, and has non-zero overlap, can we assign an outcome to the corresponding Aiomic- Therapy?
- How much of a Given-Therapy outcome can we allocate to the Aiomic-Therapy in cases where there is partial overlap between the recommended therapy and the given therapy?
- If the adoption rate was 30%, is 30% of the outcome due to the recommendation? That is to say, if only one of the drugs from the recommendation were used in the given therapy, what % is attributable to the recommendation.
- Can we create a predictive model for partial adoption of our therapies?

3 Methods

3.1 Matching score using graphs

We know the aionic therapy identifies a subnetwork of genes that is the most active for each patient. This in turn we can think as a subgraph and derive our analysis from graph theory. An important measure that comes into place is the *first Betti number* or *circuit rank*, defined as $|E| - |V| + |C|$, where E is the set of edges, V the set of vertices and C the set of connected components. In a given graph, we define a connected component as a maximal subgraph that satisfies that any two vertices are connected by a path.

In our problem at hand, by the aiomic algorithm we have identified a certain subgraph of the gene map, let's call it G. From there, the proposed therapy would try to minimize the circuit rank of the subgraph generated by subtracting the vertices and edges from G that the proposed drugs target. For a given therapy T we denote this number by B_T. In this context, it seems natural to try to quantify the change of the circut rank of the subgraph generated by subtracting the vertices and edges that correspond to the targeted genes of the actual therapy T' that the patient receives. We could propose as matching score the quantity $B_{T'}/B_T$.

3.2 Similarity measure: Jaccard similarity coefficient

Analysis of similarity of personalized cancer therapy identified by the system and the therapy clinician prescribed help us understand the gap between the "precision medicine" paradigm of cancer therapy and the therapy that patients actually received. Essentially, the presence and absence of drugs are surveyed clinical research using sequencing, imaging and other technique. Then, the Jaccard coefficients is one of the most fundamental and population similarity measures to compare such presence-absence data [**?**]. In section 3.1, we give a brief introduction of the Jaccard coefficient and we present a hypothesis test for similarity for presence-and absence data, using Jaccard coefficient based on the boottrap procedure [?] in section 3.2. The boottrap procedure is considered in order to overcome the computational burden due to the high-dimensionality. And, we claim that, if the asymptotic distribution of similarity exist and shown to be normal distribution, the present bootstrap hypothesis test can be used for any such similarity. Finial, we close the section by introduction the population hypothesis test consider using extreme value distribution [?].

3.2.1 Jaccard Similarity Coefficient for Presence-Absence Data

Due to the complexity of data set and lack of clinical knowledge, we skip the data process procedure. Consider that, each patients are given two potential therapies, one is recommended by our system and the other is the one prescribed by the clinician. After processing the data, we image that the two therapies are in the format of presence-absence vectors [?], which could be the list of the targeted gene or of the drugs recommended. Ideally, any meaningful similarity to match two therapies should display the following desirable properties:

- **Quantification:** Different similarity measures force on different types of association, such as Pearson's correlation measuring the linear relationship between variables. It is important to select the one that satisfied our request and can be used to quantify the aspect we are interested in.

- **Interpretations:** Thanks to the machine learning technique, we are able to design a specific algorithms that can be used to calculate the similarity. Unfortunately, most of them are the 'black box' computing, which is hard to do the interpretation and lack of physical meanings. Thus, the similarity measure that could be explainable is the most fit one, especially for the clinical research.

- **Statistical guarantees:** Most of similarity measures lack probability interpretations or statistical error control. And, its statistical properties, hypothesis testing, and estimation methods for p-values have been inadequately studies. Thus, a rigorous statistical test evaluating the similarity is necessary.

The one we consider here is Jaccard coefficient. Given two presence-absence vectors A and B of length m that represent two two different therapies, the Jaccard similarity coefficient is the ratio of their interaction to their union. The set A and B can be viewed as the targeted genes or the recommended drugs for the unique individuals. This quantification of overlaps allows us to quantify matched genes / drug. To explain the basic idea of Jaccard coefficient, as a toy example, suppose we two set of drugs $A = \{1, 1, 1, 1\}$ and $B = \{0, 1, 0, 0\}$. Then, the union is $A \cup B = \{0, 1, 0, 0\}$ and the intersection between the sets is $A \cap B = \{1, 1, 1, 1\}$. Jaccard coefficient can be computed based on the number of elements in the intersection set divided by the number of elements in the union set.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{1}{4}$$

Note that $0 \leq J(A, B) \leq 1$. The higher the percentage, the more similar the two sets. The formula to find the Index is:

$$\text{Jaccard Index} = (\text{the number in both sets}) \ / \ (\text{the number in either set}) * 100$$

The formula in notation is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

In Steps, that's:

- Count the number of members which are shared between both sets.
- Count the total number of members in both sets (shared and un-shared).
- Divide the number of shared members (1) by the total number of members (2).
- Multiply the number you found in (3) by 100 .

This percentage tells you how similar the two sets are.

- Two sets that share all members would be 100% similar. the closer to 100%, the more similarity (e.g. 90% is more similar than 89%).
- If they share no members, they are 0% similar.
- The midway point $-50\%-$ means that the two sets share half of the members.

3.2.2 Bootstrap Procedure for N-to-1 Clinical Trail: Patient level

From the statistic perspective, statistical hypothesis testing using this similarity coefficient provides the confidence of the result. To evaluate whether A and B are independent, a following statistical hypothesis testing is performed in the patients level:

$$H_0 : J^c(A, B) = 0, \quad H_1 : J^c(A, B) \neq 0$$

The null hypothesis H_0 is that the centered Jaccard coefficient equals zero. Note that this is equivalent to that the conventional Jaccard coefficient equals an expected value under independence. However, like most similarity coefficient, the Jaccard coefficient lacks probabilistic interpretations and statistical error control. Such problems could make results lack of generalization and confidence. In additional, another challenging presents a novel statistical problem, called the N-of-1 problem [**?**], where it is difficult to achieve statistical power. Because each patient in a trial is receiving a differing, unique therapy, these therapies are not obviously comparable.

In order to utilize the Jaccard similarity coefficient, Chung, N. C. (2019) [**?**] propose a family of methods and algorithms. As indicated in Chung's paper, an unbiased estimation of expectation and a centered Jaccard coefficient has been proposed and an exact distribution of Jaccard similarity coefficients under independence that is shown to provide accurate p-values.

Proposition 1 *(Asymptotic property)[**?**] Given A and B are independent then,*

$$\sqrt{m} J^c(A, B) \rightarrow \mathcal{N}\left(0, \sigma^2\right), \qquad \text{as } m \rightarrow \infty$$

Here, based on Chung, N. C.[**?**]'s work, we present a rigorous statistical testing to evaluate the similarity in presence-absence data, deriving statistical asymptotic properties and estimation of significance of the Jaccard coefficient.

Because the exact solution for a large m is computationally expensive or small m for lack of power, the bootstrap procedure is proposed to approximated the distribution of Jaccard coefficient. The bootstrap procedure has gained popularity for its wide applicability and statistical learning. The basic idea of bootstrap is that, by using resampling method, we could create an empirical distribution that converge to exactly distribution almost surely, And, it allows for estimation for p-values. Thus, we can access the significance of $J^c(A, B)$. In particular, resamping method with replacing A and B separately, breaks the potential dependency and make the independent assumption valid. Thus, we would be able to calculate an empirical distribution of Jaccard coefficients under the null hypothesis.

The advantages of using bootstrap procedure is (1). The expectation of Jaccard coefficient can be estimated directly from resampled vectors $A^\star$ and $B^\star$; (2). Each iteration provides randomness, which helps avoid a bias related to using an estimated expectation based only on observation. (3). Under the setting: N-to-1 clinical trails, we could avoid the request of large samples and be able to performing the statistical hypothesis testing for each unique patient.

Algorithm 1: Bootstrap Procedure for Jaccard coefficient

```
Input: Two binary therapy $A$ and $B$;
1. Calculate a centered Jaccard coefficient;
while $k = 0, 1, \cdots$ do
    Resample with replacemebt $A$ and $B$, resulting in $A^\star$ and $B^\star$;
    Calculated boottrap null coefficients
end
Compute the $p$-value by
    $p\text{-value} = \frac{\mathbf{1}\{|t_b^*| \geq |t|; b = 1, \ldots, B\}}{B}$
```

3.2.3 Population Hypothesis Testing

In this section, instead of considering patient level analysis, we focus on the hypothesis test for group of patients. Let assume we have a vector of Jaccard coefficient, J. we consider to using the minimum extreme value distribution to evaluate the significance of Jaccard coefficient borrow the idea of Rahman et al. (2014). Rahman et al. (2014)[**?**] proposes a method to compute a p-value of a Jaccard coefficient using an extreme value distribution.

For the statistic hypothesis, we need to find a statistic that characterize the samples and can be used of testing. We are often interested in extreme values of a parameter, like minimum Jaccard coefficient in our study and minimum strength, minimum force, minimum net income in a stock, because they are the values that determine whether a system will potentially fail or the minimum benefit that guaranteed. For example: minimum net income in a stock - it must be arranged to be at least greater than zero to prevent the cost; minimum risk of prescribed therapy that ensure patients is in safe side; modeling the extremes of meteorological events is shown to be necessary since these cause the greatest impact. It worth noting that the extreme value distributions are asymptotic results, meaning that the probability distribution of the minimum of a set of m independent values drawn from some distribution approaches the extreme value distributions only as n approaches infinity.

$$\text{Probability Density Function: } f(x) = \left(\frac{1}{b}\right) \exp\left(-\frac{x-a}{b}\right) \exp\left[-\exp\left(-\frac{x-a}{b}\right)\right]$$

Measuring statistical significance of the hits: The significance of the hits returned from the database can be inferred from the p-values derived from the z scores of the similarity.

The mean (μ) and s.d. (σ) of the similarity scores are used to define the z score, $z = (J - \mu)/\sigma$. For the purpose of calculating the p-value, only hits with $J > 0$ are considered. The p-value is derived from the z score using an extreme value distribution. And the p-value is calculated as below:

$$P = 1 - \exp\left(-e^{-z\pi/\sqrt{6}-\Gamma'(1)}\right), \text{ where the Euler-Mascheroni constant } \Gamma'(1) \approx 0.577215665.$$

3.3 Statistical Analysis

In this section, we give the pipeline of the statistical analysis. Note that we only provide the general framework and some necessary adjustment will be made based on the data set received. The primary outcome is to examine the impact of the matching score on the clinical measure, such as survival time. Although, it is a indirect evidence, it provides the idea of accuracy of the the system and could help the improvement. Before go that deeper, we first start with some exploratory analysis which can help locate the population and give the general picture of the samples.

Preliminary Screening: Prior to conducting the exploratory factor analysis, preliminary screening will be conducted. Data will be first screened for inclusion/exclusion, consent, end of study completed, missing data. All the inclusive and exclusive condition need to be satisfied and the data of participants can only be used after consent.

Descriptive Statistics: Patients' demographic information and clinical characteristics will be examined with descriptive statistics, using frequency (percentage) for categorical variables and mean (standard derivation), median (interquartile range) and range for continuous variables as appropriated.

Survival Analysis: Survival analysis will be used to study the association between the similarity of therapies and clinical measurement (time-to-death). Logrank test, Wilcoxon test, Fleming test and Kaplan-Meier analysis will be used to visualize the estimated probability of survival given specific time point and compare groups of patients (patients with similarity $\leq \alpha$ vs. the rest). p-values ≤ 0.05 are considered significant. And p-values-values will be adjusted for multi-comparison if needed.

In order to adjusting the validation cased by patients, mixed effect univariate and multivariate cox proportional hazards regression models will be used to estimate the hazard ratio of similarity coefficient to the survival time. The proportional hazards assumption will be tested by assessing Schoenfeld residuals and by plotting the negative logarithm of the estimated survivor function against the log time using log plots.

3.3.1 Mixed effects Cox Regression Models

Mixed effects cox regression models are used to model survival data when there are repeated measures on an individual or some other reason to have both fixed and random effects. The mixed effect cox regression model fits the model

$$\lambda(t) = \lambda_0(t)e^{X\beta+Zb}, \quad \text{where } b \sim G(0, \Sigma(\theta))$$

where λ_0 is the baseline hazard function, X and Z are the design matrices for the fixed and random effects, respectively, β is the vector of fixed-effects coefficients and b is the vector of random effects coefficients. The random effects distribution G is modeled as Gaussian with mean zero and a variance matrix Σ, which in turn depends a vector of parameters θ.

The main idea of mixed effects cox regression models is that they make specific assumptions about the variation in observations attributable to variation within a subject and to variation among subjects. Under our setting, we are consider the variation is coming from the uniqueness of the patients.

3.3.2 Mixed Effect Model

To make the explain the mixed-effect in a easy way, we break the cox regression model into two part, linear model and link function, where we have linear model as $Y = X\beta + Zb$ with link function: $g(t) = \lambda_0(t)e^Y$. It's not hard to see the cox regression is kind like perform a map (link function) on a linear model.

We use a simple notation for convenient and ignore the link function for now. let Y_{ij} denote the response of subject $i, i = 1, \ldots, n$ at time $X_{ij}, j = 1, \ldots, n_i$ and $\beta_{i0} + \beta_{i1}X_{ij}$ denote the line that characterizes the observation path for i. Note that each subject has an individual-specific intercept and slope. Note that

- The within-subject variation is seen as the deviation between individual observations, Y_{ij}, and the individual linear trajectory, that is $Y_{ij} - (\beta_{i0} + \beta_{i1}X_{ij})$.

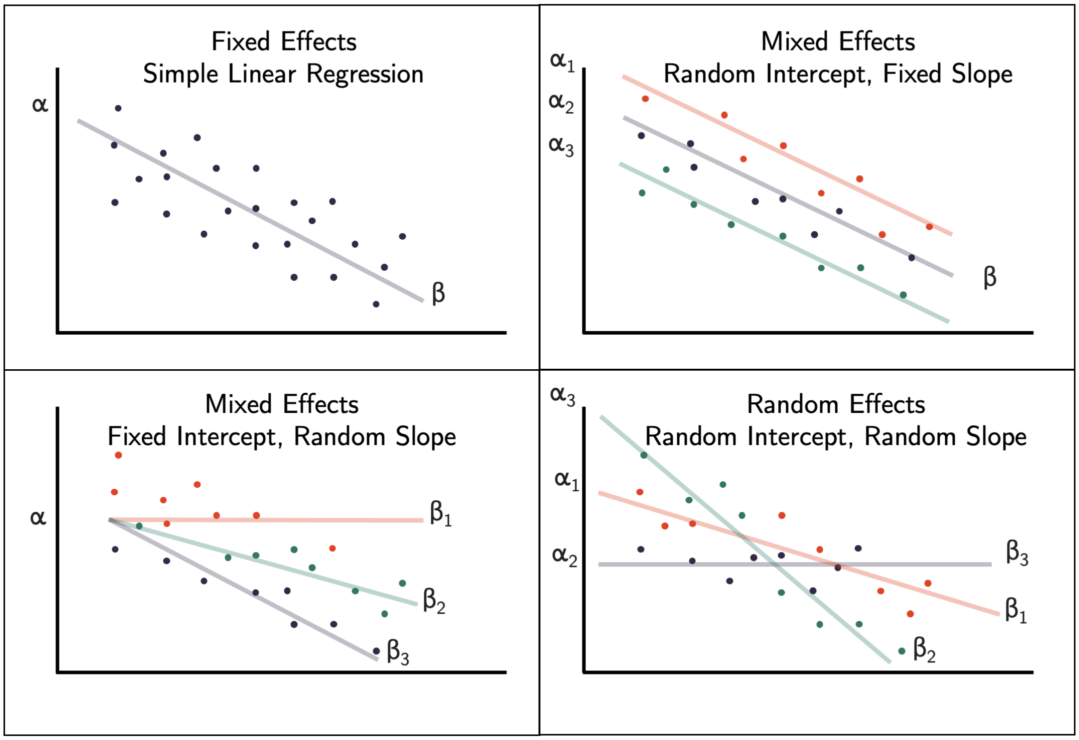
$$E\left(Y_{ij} \mid \beta_i\right) = \beta_{i,0} + \beta_{i,1}X_{ij}, \quad Y_{ij} = \beta_{i,0} + \beta_{i,1}X_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

- The between-subject variation is represented by the variation among the intercepts, $\operatorname{var}\left(\beta_{i0}\right)$ and the variation among subject in the slopes $\operatorname{var}\left(\beta_{i1}\right)$.

$$\begin{pmatrix} \beta_{i,0} \\ \beta_{i,1} \end{pmatrix} \sim N\left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix}\right]$$

where D is the variance-covariance matrix of the random effects, with $D_{00} = \operatorname{var}\left(b_{i,0}\right)$ and $D_{11} = \operatorname{var}\left(b_{i,1}\right)$

From the following figure, we can say that (A) A randomintercepts model where the outcome variable Y_{ij} is a function of predictor X_{ij}, with a random intercept for study ID. Because all individuals have been constrained to have a common slope for predictor X, their regression lines are parallel. Solid lines are the regression lines fitted to the data. Point colour corresponds to study ID of the data point. The black line represents the global mean value of the distribution of random effects. (B) A random intercepts and random slopes model, where both intercepts and slopes are permitted to vary by group. Random slope models give the model far more flexibility to fit the data, but require a lot more data to obtain accurate estimates of separate slopes for each group.



3.3.3 Multivariable-Level Analysis

Mixed Effects Cox Regression Models will be used in multivariable-level analysis in order to adjust patients' individual effect and to quantify the effect of selected variables when they cooperate by presenting the coefficient (standard error), 95 % confidence Interval and P-value. The variables included in the multivariable analysis was selected if they were statistically significant on the univariable level analysis and considered clinically significant by the research team.

Multivariable-Level Analysis help to identify the risk factor and quantify their impact when they work together. All the risk factor that shows significant could be consider as the main factor to be used in the system to predict the gene and help increase the accuracy. Besides, several advantages can be used to help identify the factor, such as variable selection technique.

4 Conclusion and Limitations

In this paper, we present a general frame for analysis the data. Specially, we propose to using Jacard coefficient as well as bootstrap procedure for statistical testing. Although, bootstrap has its advantage of lower computation cost, unfortunately, when the size of samples are extreme small, there is no statistical guarantee and the result is obvious lack of reliability. Such restriction should be take into consideration when it comes to the real-data analysis and other method can be consider to address this issue. Another changeling is that potential problem with this score based on graph theory: We know that drugs in general target more than only one gene. Potentially, this gene could very well be outside of our principal subgraph G, but could have interactions. How can we incorporate this into our matching score? We leave these problems for later work

Acknowledgements This works was supported by the PIMS Math industry work shop and we would like to express our appreciations to our academic mentor, Dr.Sang, andindustrial mentors, Ali Hashemi and Giannoula Lakka, for advice and assistance with their knowledge on this project.

Math INDUSTRY

CITY OF WINNIPEG ICB: MODELLING MOSQUITO POPULATIONS

ANDRII ARMAN, JONATHAN GALLAGHER, AND AIDIN ZAHERPARANDAZ

Abstract. The city of Winnipeg's Insect Control Branch (ICB) is interested in predictive models of mosquito population based on environmental data. In particular, ICB determines whether to spray or not a given region of the city based on a mosquito trap count.

In this project we collect and aggregate data provided by the city of Winnipeg and external sources, determine key weather factors that influence mosquito population, and provide a model (Random Forest Classifier) that determines whether mosquito count in a given region of a city is larger than a given threshold. As a part of data collection we develop a computer vision tool for precipitation data extraction from satellite images.

1. Introduction

The City of Winnipeg Insect Control Branch (ICB) provides services to public to control mosquito population. Mosquito control program includes helicopter larviciding program, ground larviciding program, and monitoring adult nuisance mosquitoes (New Jersey Light Traps).

Main challenges for ICB are: predictive modelling of rainfall/soil moisture, and predictive modelling of larval and adult mosquito population. In particular, a key factor for ICB on deciding whether to spray in a given location of a city is whether a mosquito count in a given location exceeds 25.

Our main contributions are in the following three directions.

To start with, a large part of our work is devoted to data collection and data aggregation. For instance, we use weather data from external source visualcrossing.com, where only average metrics were given for a city of Winnipeg. In a pursuit of more localised weather information (in particular precipitation), a computer vision tool is developed that can be used for weather information extraction from satellite images, weather maps, etc. We discuss data collection and a concept of data silo in Section 2.

Second, we determine key factors that influence mosquito population. For instance minimal temperature yesterday happens to be the most significant weather factor for today's mosquito population. The analysis is done via important features of linear regression model in Section 3.

Finally, we obtain a predictor of whether a $trap\_loc[date] > 25$, where $trap\_loc[date]$ is a mosquito count in a given location on a given date (ICB makes a decision to spray a given location if $trap\_loc[date] > 25$). Our predictor of $trap\_loc > 25$ is a Random Forest Classifier that has good accuracy ($> 89\%$ for all regions) and precision ($> 84\%$ for all but three locations). Description of this predictor is given is Section 4.

2. Towards distributed and accessible city-data

2.1. **Data silo.** The concept of a *data silo* refers to a situation where organisational data is not collaboratively accessible. Data silos occur for different reasons; for example, the data could be held by different groups within an organisation (perhaps unknown to each other), or stored in seemingly incomparable formats. Note: often data silos refer to informational systems that should be linked together, and often excludes insulated data warehouses that are necessarily incompatible – in other words, linking data between silos would not, in theory, break the principle of least access. Data silos present a significant obstacle to large institutions (governments, large corporations) [3] because they lead to

(1) replicated efforts in data analysis;
(2) unawareness of the available data;
(3) overly specialised and non-transferable knowledge;
(4) a lack of shared incentives and objectives [2].

Breaking down data silos can lead to more shared incentives as well as faster and more complete analysis [3].

An additional obstacle for city-level data is that often the data is collected over multiple decades and is potentially dangerous or disruptive to reformat. A general solution to this sort of age-layered problem is known as continual improvement [1]. In the remainder of this section, we describe our efforts to start continuous transformation of the disparate and sometimes lacking data sets using a lightweight approach to what is known as data virtualisation.

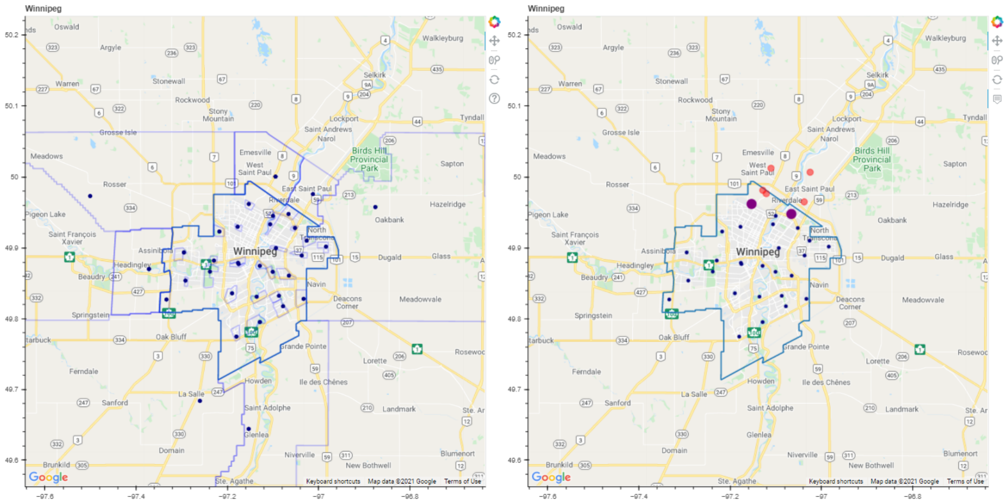
Data virtualisation refers to techniques that allow amalgamating data from disparate, heterogeneous, and discordant sources. Our data sources include:

(1) existing data sources owned by the ICB;
(2) data sources owned by other departments are added;
(3) relevant data sources owned by outside departments (e.g. the European Space Agency and visualcrossing.com) are added.

2.2. **Amalgamating and harmonising data.** The following data sources are amalgamated into a single data-frame:

- A database of daily mosquito trap counts for 2015-2021;
- A file containing mosquito trap locations (36 locations);
- 311 City requests (mosquito complaints) for 2014-2021;
- A database of helicopter larviciding data for 2020-2021;
- A database of ground larviciding data for 2003-2021;
- A listing of ground larviciding locations;
- Weather data for 2015-2021 (externally sourced from visualcrossing.com).

After a simple amalgamation of the databases, we added spatiotemporal dimension to the dataframe. For example, daily mosquito trap is indexed by date with columns being trap locations: 28 locations $NW_1, ..., SE_7$ and 8 locations outside city $AA, ..., HH$. This data is then cross-referenced with the mosquito trap location data. Mosquito trap locations is a JSON file that for each trap name provides a polygon in which trap is located. See Figure 1a, blue dots represent the center of a corresponding trap polygon. This means that within the



(A) Trap location areas and approximate locations

(B) Helicopter spraying and closest traps

FIGURE 1. Visualisation tool use example (for 2021/5/27)

amalgamated dataframe, mosquito trap counts can be looked up by date and location within the city, and also that date-location may be looked up to determine mosquito trap counts.

Mosquito complaints are indexed by date, and contain the location of a complaint. While we did not explicitly extend the dataframe with 311 data indexed by date, date-location indices can easily be extended to include 311 data.

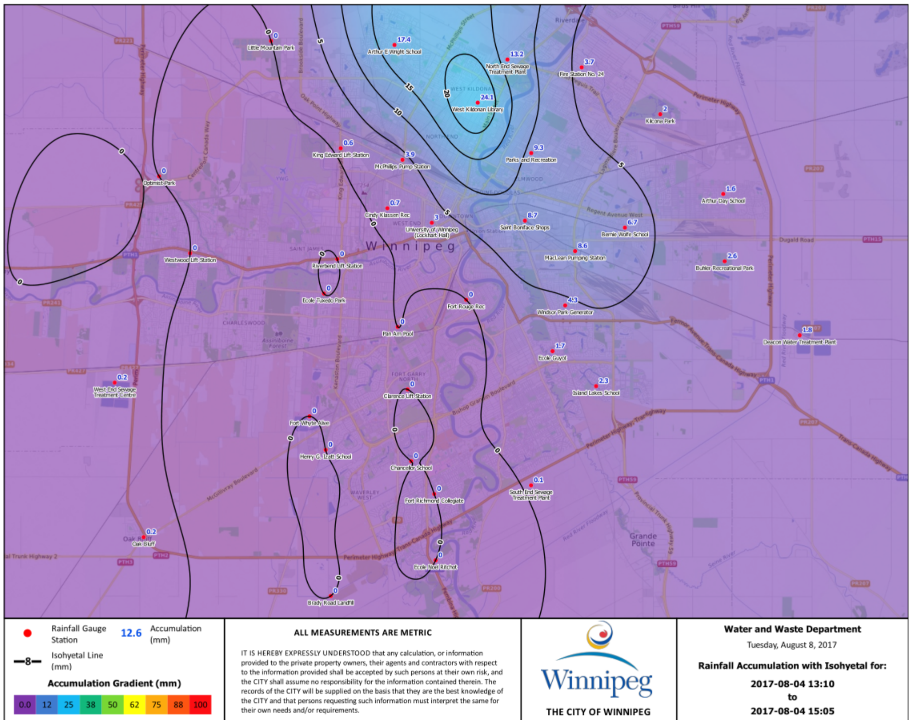
Helicopter larviciding data is indexed by datetime and contains, principally, a polygon indicating where spraying occurred. We replaced spraying polygons with their centers, and for each spraying we found the closest trap to the spraying location. This data was categorical; for each day and each trap we recorded 'yes' or 'no' based on whether or not spraying occurred within a bounding box containing the trap. For example, on Figure 1b spraying locations (red dots) on a given day (2021/5/27) are presented; the closest traps (NW6, NE1) are highlighted in purple.

Ground larviciding data is similar design to helicopter larviciding data. We received ground larviciding data too close to the end of the project to incorporate it, but the same procedure as used for helicopter larviciding data could be used to ground larviciding as well.

Weather data includes temperature (average, min, max), precipitation, wind, etc (obtained from visualcrossing.com). This data contains only city averages, and is added to the dataframe by date, as each location is taken to have the same data. Initial analysis indicates that the most important weather features are temperature, precipitation and cloud cover.

To create Figures 1a and 1b we use the Bokeh Python library for creating interactive maps. The resultant maps and the ease of overlaying additional data on demand could be developed further into a powerful and flexible visualisation tool.

2.3. **More accurate rainfall data.** Since precipitation data is identified as a significant data point for mosquito population modelling, the accuracy of precipitation data is important. The data from visualcrossing.com provides only city averages, and some accuracy issues have been identified. The City of Winnipeg, Water and Waste department maintains maps of rainfall as collected by multiple sensors spread throughout the city (see Figure 2).



ALL MEASUREMENTS ARE METRIC

IT IS HEREBY EXPRESSLY UNDERSTOOD that any calculation, or information provided to the private property owners, their agents and contractors with respect to the information provided shall be accepted by such persons at their own risk, and the CITY shall assume no responsibility for the information contained therein. The records of the CITY will be supplied on the basis that they are the best knowledge of the CITY and that persons requesting such information must interpret the same for their own needs and/or requirements.

Water and Waste Department
Tuesday, August 8, 2017
Rainfall Accumulation with Isohyetal for:
2017-08-04 13:10
to
2017-08-04 15:05

FIGURE 2. An example of rainfall map from Water and Waste department

The raw data of the rainfall amounts is not available, leading to a silo problem (this data cannot be directly integrated). To circumvent this issue, we created a tool that loads images in pdf format directly from the Water and Waste department website, converts them to png format, and uses machine learning to to reverse-engineer the rainfall amount.

The heart of determining the rainfall amounts from an image amounts to subdividing the image into a grid of smaller squares (see Figure 3 for an example); this is a configurable option of the tool. Then we remove certain features that have nothing to do with rainfall amounts: this includes the description and information block at the bottom of the image and the isohedral lines laid over the image. To perform this, we tie rainfall location amounts to rainfall gauge stations only; thus, only squares in the grid containing a rainfall gauge station are counted (the rest of the map is extrapolated anyway). To remove the isohedral lines, the colour black is consistently used for drawing the lines and never used for indicating rainfall amount. So we simply omit certain ranges in the RGB spectrum. We use the Open-CV https://opencv.org/ library for computer-vision to transform an image into a 2-dimensional array of RGB-values. We use the SciPy library https://www.scipy.org/scipylib/ to apply the k-means algorithm to each square in the grid and determine the k most dominant colours in that grid. For example, for the grid square in Figure 3 and $k = 2$ we get the two most dominant colours as a light-purple and a light-blue.

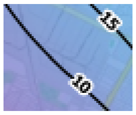


Figure 3. An example of a small square obtained from Figure 2

We then select the most dominant colour by applying a k-clustering algorithm (also from the SciPy library) to the k-means, and selecting the colour that has the largest cluster around its centre. This way we obtain the colour value that we associate to the current grid square. As the images produced by the Water and Waste department do not use a consistent colour scheme for rainfall amounts, we allow passing the colour encoding as a parameter, and then encode the different colour schemes used to allow determining rainfall amounts from a colour. For a piecewise linear colour gradients, we include a utility to automatically invert the colour scheme. Finally, given a mapping of dates to colour schemes, we process a range of dates to produce rainfall amounts over the city by location and date.

Since this data is now inherently made spatiotemporal, it may be readily added to our dataframe, though we haven't done this yet. It is worth noting that in the case of archived data, where the source data is not kept and only artefacts (such as rainfall maps) are kept, the technique of detecting specific features by combining the OpenCV library and a tool such as k-means can be reused. This technique may also be extended for non-specific features (such as shape of rainfall density over the city) by using neural networks.

2.4. **Moisture data.** While rainfall amounts are indicative of mosquito behaviour, soil moisture is seen to be more important for directly modelling mosquito populations. Rainfall does not directly translate to soil-moisture, as dry soil has different absorption properties compared to already moist soil.

One of the European Space Agency's (ESA) public projects is satellite data for soil moisture (and oceaninc salinity) https://earth.esa.int/eogateway/missions/smos. The soil-moisture data is provided in a general format though not immediately accessible for automated processing (it's accessible by ftps, and easy to access manually). To support obtaining soil-moisture data tied to geographic locations, we created a tool to automatically access the data and extract the soil-moisture data required. The only thing that is needed to run the tool is an ESA account (the sign-up form can be currently found https://eoiam-idp.eo.esa.int), and the desired start and end dates for obtaining the data. We automatically login to the ESA SMOS repository, access "level 2" science products (this includes the soil-moisture data), and download the data for the selected date range. The SMOS products contain soil-moisture data for nearly the entire planet, and includes earth-explorer data files that create global visualisations of soil-moisture (see Figure 4)

However, the SMOS data also includes raw data. We access this by downloading the NetCDF https://www.unidata.ucar.edu/software/netcdf/ formatted data, and then automatically filtering and extracting the soil moisture amounts by location.

This also makes the SMOS data fit consistently with our spatiotemporal dataframe. While we have not integrated the tool to update the dataframe, with the tool itself complete, this could be readily completed. As the processing we perform on this data is involved, we can

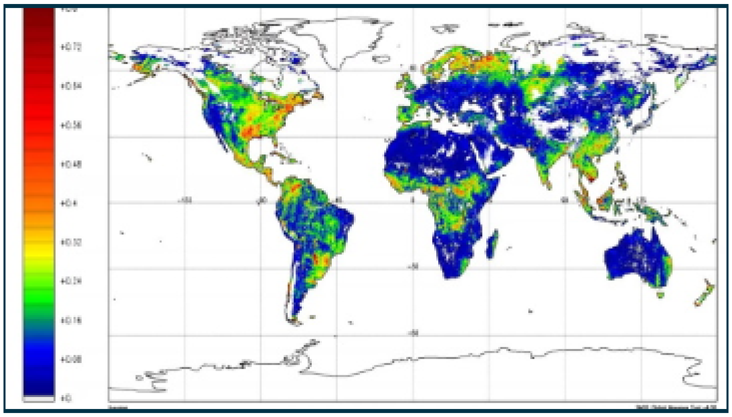


Figure 4. An example of soil moisture map from SMOS

also make the virtualized data faster to use by locally caching the SMOS data and preferring cached data if it exists. Thus, after a first run, accessing this data should be fast and reliable.

Much of the processing required is already performed by the SMOS project, including neural networks for determining soil moisture from the raw satellite image data. This could be seen as disadvantageous if one desires using a neural network that allows for more accurate moisture amounts to be determined using more locally available contextual information. However, the SMOS project does provide the raw satellite data in their "level 1" products section, and integrating these data points for more fine-grained analysis is feasible.

3. Identifying the most significant features for mosquito population

In order to determine the relationship between the acquired weather data, and the available data from helicopter spraying program and mosquito count from traps in 28 parts of the city, we have made predictions based on a linear regression model in three phases.

3.1. A sight from regression analysis. Once we had pre-processed all the gathered data from our datasets and the daily weather information, we ran a linear regression model to see the effect of current day weather information including precipitation, temperature, cloud cover, humidity, and wind speed on average trapped mosquito count across all over the city for each year from 2015 to 2021, as well as all years to determine the significance of each feature.

In order to check the importance of each feature in our regression model and to prevent multicolinear features, we took advantage of a statistical tool called "Variance Inflation Factor" or VIF, which provides an index to reflect the increase of variance of an estimated regression coefficient due to multicolinearity. We have dropped the features with VIF > 10 as it may not add to the descriptive power of our model.

Through this step, we essentially drop the most correlated features and identify most relevant features as the minimum temperature, cloud cover and precipitation.

This initial linear regression model was a foundation for developing further regression models through our analysis; in spite of the fact that it was not considerably successful at making highly accurate predictions for the average mosquito count in the city.

3.2. **Seven-days window of features.** After developing our first working linear regression model, we decided to consider a window of seven days of weather conditions to predict mosquito count. We also localised our model so that we could make predictions for each individual location of the city instead of the average mosquito count city wide. In this model we also included the helicopter spraying of chemicals for the past seven days.

These considerations led us to build a model with 26 features including helicopter spraying, precipitation, cloud cover, and minimum temperature for the past seven days.

The resulting model is able to explain more than 90% of the variability of the data with a better accuracy for all regions. However, there was a risk of overfitting, slightly higher VIFs for some features, and auto-correlation of features during the seven days period are serious drawbacks of this model.

For example, the model for the region NW1 in the year 2020 is making predictions for mosquito count using the above-mentioned features and provides an explanatory power of 96%. The evaluation of the targets and predictions using the model with 26 features is shown in Figure 5 to demonstrate a visual interpretation of prediction accuracy and it helps us in upgrading our model in making better predictions.

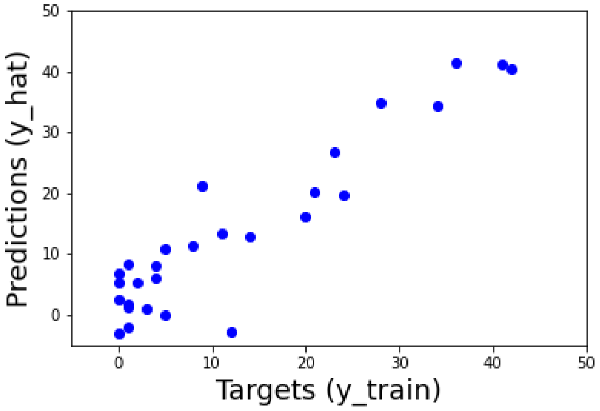
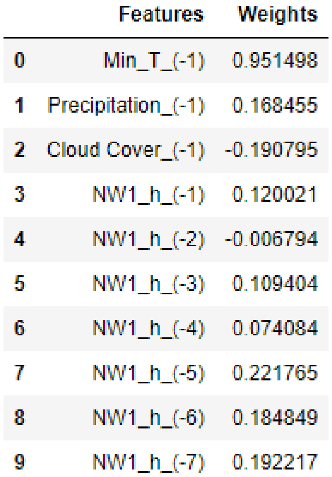


FIGURE 5. Targets vs. Predictions for mosquito count - distribution around 45° line as an illustration of model accuracy. Second linear regression model with 26 features and 96% explanatory power for NW1-2020.

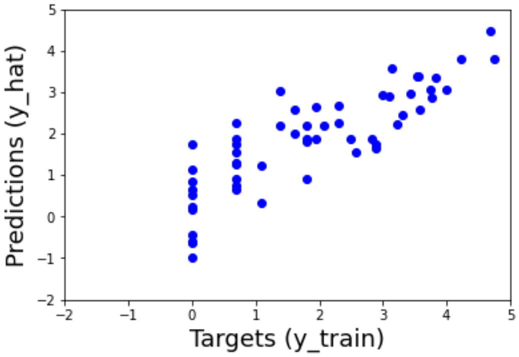
3.3. **Picking the best describing and most significant features.** In the last step of improving our linear regression model we narrow down our focus on only ten most important features including the minimum temperature, cloud cover, and precipitation for the previous day, as well as helicopter spraying in the past seven days. We also consider log of mosquito count as a target variable. These features have the least VIF scores, representing their least multicolinearity, and provide us with an improved balance of predictive power and explanatory power.

As per our analysis, we obtain an explanatory power of more than 70% for most regions using this model. While this model yields to a relatively strong explanatory capability, it had pretty low prediction accuracy. In each region of the city as this model has shown an accuracy of 30% to 40% in making predictions. However, this linear regression model robustly shows the relationship between the best describing features and the target and makes it available for us to build stronger models.



	Features	Weights
0	Min_T_(-1)	0.951498
1	Precipitation_(-1)	0.168455
2	Cloud Cover_(-1)	-0.190795
3	NW1_h_(-1)	0.120021
4	NW1_h_(-2)	-0.006794
5	NW1_h_(-3)	0.109404
6	NW1_h_(-4)	0.074084
7	NW1_h_(-5)	0.221765
8	NW1_h_(-6)	0.184849
9	NW1_h_(-7)	0.192217

(A) Coefficients table



(B) Targets vs. Predictions - log mosquito count shown - distribution around 45° line as an illustration of model accuracy.

FIGURE 6. Final linear regression model with 10 most significant features and 71% explanatory power for NW1-2020.

In Figure 6 we illustrate our results using this model for station NW1 in the year 2020 where we obtain a 71% of explanatory power with the provided ten features and their corresponding coefficients in Figure 6a, but making accurate predictions is what we need to improve for this case as it is shown in Figure 6b.

4. Random Forest classifier for windowed, mosquito threshold predictions

We use Random Forest Classifier (RFC) to obtain a predictor for $trap\_loc[date] > 25$ for each of 28 trap locations.

RFC is a classifier that is based on decision tress. In a decision tree classifier, we want to decide whether $trap\_loc[date] > 25$, based on a series of simple yes/no questions. For example, we may ask if minimum temperature exceeds $15°C$, or if there was more than 1in of precipitation to determine if mosquito trap count exceeds 25. The problem with a decision tree classifiers is that they tend to overfit data by having large depth. Hence decision trees will likely perform well on training data, but might underperform on testing set.

One way to prevent overfit, is to consider a random forest: an ensemble of tress that use random subset of features and a random subset of training data. An oversimplified view of random forest is that we have many 'expert' trees that can fit a given data really well, but to each 'expert' we disclose only some portion of features and some portion of data. As a final prediction we take a majority of all of our many 'expert' decisions.

We use weather data (minimal temperature, precipitation, humidity) and trap count data as initial data for a given location. We use $\sim$ 1000 data points (years 2015-2021). We use back-filling for missing entries in trap count data.

Our classifier has 24 features as input: minimal temperature, precipitation, humidity and *trap_loc* over 6 days window (from date-2 to date-7), and target is a boolean $trap\_loc[date] > 25$.

We use 20 trees in each forest and 5 features per each tree. The split between training and test data in 70%/30%. We used RandomForestClassifier provided in sklearn Python library.

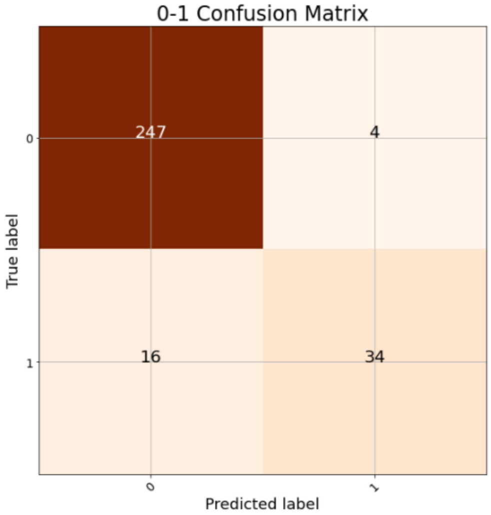


Figure 7. Confusion matrix for NW1 location

For illustration purposes, the confusion matrix of performance of RFC on testing data for NW1 region is presented in Figure 7. This matrix illustrates performance of our model on test data, with off-diagonal numbers indicating misclassified instances (true-negative and false-positive). We use accuracy and precision to estimate RFC performance. From the matrix in Figure 7 we deduce that RFC for NW1 location has accuracy 0.93 and precision 0.89.

loc	acc	prec	loc	acc	prec	loc	acc	prec	loc	acc	prec
NW1	0.93	0.89	NE1	0.89	0.79	SW1	0.96	1.00	SE1	0.96	1.00
NW2	0.95	0.95	NE2	0.92	0.89	SW2	0.90	0.84	SE2	0.94	0.90
NW3	0.97	1.00	NE3	0.95	1.00	SW3	0.91	0.93	SE3	0.94	1.00
NW4	0.90	0.84	NE4	0.99	1.00	SW4	0.94	1.00	SE4	0.93	0.84
NW5	0.96	1.00	NE5	0.95	1.00	SW5	0.98	0.00	SE5	0.92	1.00
NW6	0.96	1.00	NE6	0.92	1.00	SW6	0.95	0.96	SE6	0.99	1.00
NW7	0.93	0.90	NE7	0.94	1.00	SW7	0.99	NA	SE7	0.93	1.00

Table 1. Accuracy and precision of RFC for different locations

In Table 1 we list accuracy and precision of RFC for all 28 trap locations. We note that for SW5 and SW7 locations there were less than 1% of data points with $trap\_loc[date] > 25$

(for other regions this value is above 10%), hence high accuracy, but low precision for these locations. For these two locations we suggest using lower thresholds (15, or 10) for *trap_loc*.

For a future work, our classifier can be modified to obtain a predictor (i.e. use classifier for a tens digit of mosquito count) for mosquito population. Additionally, we do not use helicopter spraying data, as the corresponding dataset is to small, however ground treatment data is extensive and could further improve performance of the model.

5. Next steps and possible future studies

Within our analysis on predictive modelling of mosquito population in Winnipeg, we tried a number of ways to approach this problem from different perspectives and our team has noticed that there might be a possibility to work on this project further.

City data was slightly disorganised, for instance we were pulling weather data from external sources, instead of using internal data (for example, Water and Waste department has precipitation data, but we were not able to obtain it). Even for the datasets available, the fields did not always match and we had to do some data patching. We identified critical steps to design a more efficient city data portal with more harmonised data frames. For instance, used for city portal, our methods allow to request all available data within a certain radius of a given geolocation.

Also, having a control study (with no spraying) on mosquito larval development in Manitoba can potentially make an improvement in our mosquito population models. Such a "pure" model for a mosquito population growth can help us with filling out the gaps and missing values in our data more accurately.

By incorporating time series analysis we can make predictions based on mosquito population patterns and eliminate the auto-correlation of the features and targets, including weather data and trapped mosquito count.

Gaussian Process Regression would be another useful approach for this project that allows a regression to take into account prior knowledge and perceived conditional probabilities to describe the future and its uncertainty.

6. Acknowledgements

To Allen, Kristine and Ruth: thank you for organising this workshop. Thanks to Julien Arino for providing useful references. And a special thanks to the ICB branch staff: Scott, David, Ken and Jennifer.

References

[1] K. Fryer, J. Antony, and A. Douglas, *Critical success of continuous improvement in the public sector: a literature review and some key findings*, TQM **19** (2007), no. 5, 497–517. ↑2.1

[2] B. Gleason and M. Rozo, *The silo mentality: How to break down the barriers*, Forbes (2013Oct). ↑4

[3] A. study group on functional organization, *Organizational renewal: Tearing down the functional silos*, Association for Manufacturing Excellence, 1988. ↑2.1, 2.1

Email address: andrew0arman@gmail.com

Email address: jonathan@infinitylab.io

Email address: aidinzp@gmail.com

Math INDUSTRY

ANALYTICS AND VISUALIZATION OF POLITICAL DATA

MITCH HASLEHURST, ANIKET JOSHI, GUOJUN MA, SAM SIMON, SHEN-NING TUNG

1. Introduction

In this report, we analyze the voting records and visualize the behaviour of politicians from both Canada and the United States of America. The data analyzed here includes details of the legislators, such as education level, place, active year in politics, the topics of bills passed by the legislators, historical voting record, and so forth. The goal is to build metrics to quantitatively measure the performance of politicians using these data. One obvious and commonly employed approach to measuring performance is via approval rating, but we seek an approach which is more rigorous. Alternatively (as a starting point), the quality of a politician can be measured by the number of bills they sponsored, or the number of bills passed as laws align with a their political belief. It is reasonable to assume when a politician vote "yes" to a bill, it means the bill aligns with their political belief, or vice versa.

The idea of measuring a politician's performance is similar to sport analytics, which has become a crucial part of almost every major sport in recent decades. For example, in the current NBA, the team management often analyzes the player's strengths and weaknesses by looking at the historical data such as where the players most often take shots in the field, points per game, assists per game, re-bounces per game, and so forth. There is little exaggeration in claiming that a given player's market worth is entirely determined by their statistics. In recent years, many different kind of statistical methods have been applied to politics, such as using the poll before a major election to predict the result (perhaps the most well-known one), conducting a survey to obtain approval ratings, analyze politicians' speech to find topic distribution, and many more.

We approach the search for such methods of measurement through several lenses. In the first section we examine how often, in Canadian Parliament, the outcome of a vote fares in favour of a given legislator. In the third and fourth sections the number of bills sponsored by both Canadian and American legislators is tracked, along with the topics of bills. In the fifth section we delve further into bills sponsored, examining the topics closely and comparing their global versus national impact. Lastly, we discuss the issue of bipartisanship in the United States. To achieve this we construct a graph to describe the interaction between politicians based on their political party.

All data used in this report is courtesy of IOTO International, and was cleaned and analyzed using Python.

Received by the editors September 10, 2021.

2. Canadian Party Leaders

One thing that we can measure for Canadian politicians in Parliament is how often a vote goes in favour of how they vote. That is, if they vote yes on the bill and it passes or they vote no on the bill and it doesn't pass. This could be comparable to scoring in various sports or for example hits in baseball. In this section we look at this voting record data for Canada's current three majority party leaders, i.e. Justin Trudeau, Jagmeet Singh, and Erin O'Toole. We compare them against each other for different parliamentary sessions and also look at individual performance over different parliamentary sessions.

We may consider each party as a "team," each Parliamentary session as a "season" and each party leader as "team captain." In this regard, there is a natural correlation between the "team's performance" and the "captain's performance". That is, when your team is doing well in the season, i.e. the Prime Minister is a member of your party, then your team captain is also doing well. Similarly if your team is not doing well that season, i.e. the Prime Minister is not a member of your party, then your team captain also does not have the best performance. In the following figures, "yy" means "voted yes and passed", "yn" means "voted yes and didn't pass", "ny" means "voted no and passed", and "nn" means "voted no and didn't pass". Thus, the sections of blue and red are those that correspond to a vote going in a politicians favour.

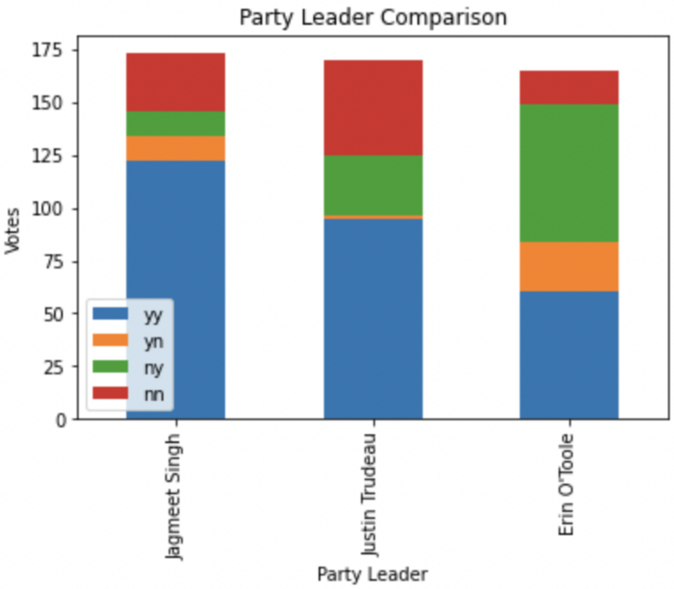


Figure 1

In Figure 1 we compare the voting record for the three party leaders over the parliamentary sessions 43-1 and 43-2. Of note, there is only one time that Justin Trudeau votes "yea" for a bill that didn't pass. Conversely, Erin O'Toole had the largest amount of bills that didn't go in his favour. Both of these trends are expected to some extent. Justin Trudeau is the Prime Minister and leader of the

Liberal party, so it is natural that many of the bills go in his voting favour. However, Erin O'Toole is the leader of the Conservative party which has historically differed with the Liberal party on many issues.

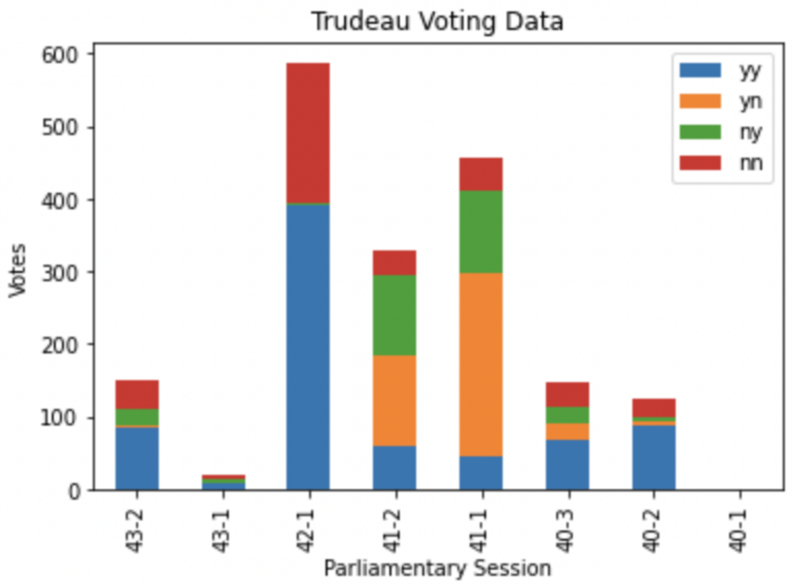


Figure 2

Figure 2 shows Trudeau's voting performance from Parliamentary session 40-1 to 43-2. We see that in the 41st parliament, most of the votes didn't go in a favourable way for him. Contrast this to the Parliament immediately after, where almost all of them did. Naturally, this is also expected to some extent as the Prime Minister transitioned from Harper (Conservative) to Trudeau himself (Liberal).

Figure 3 shows O'Toole's voting record. Comparing to Trudeau's we see a mirrored picture in the sense that the sessions where O'Toole performs well are where Trudeau performs poorly, and the reverse also holds. We also note that during the 41-1 and 41-2 Parliamentary sessions that almost every bill went in O'Tooles favor. In fact, only 3 votes out of a total of 602 votes when against his own vote. This was during Harper's term as Prime Minister, so one might naturally question the similarities between the two. However, when comparing them directly, we see that they voted differently on 131 bills out of the 591 that they both voted on. Therefore, O'Toole more closely resembles the political policies but in place through Parliament during the 41st session rather than Harper's personal voting record.

We remark that this method of comparison and "scoring" is rather naïve. Although the Prime Minister is from one of the main parties, the composition of the Parliament is also important to consider.

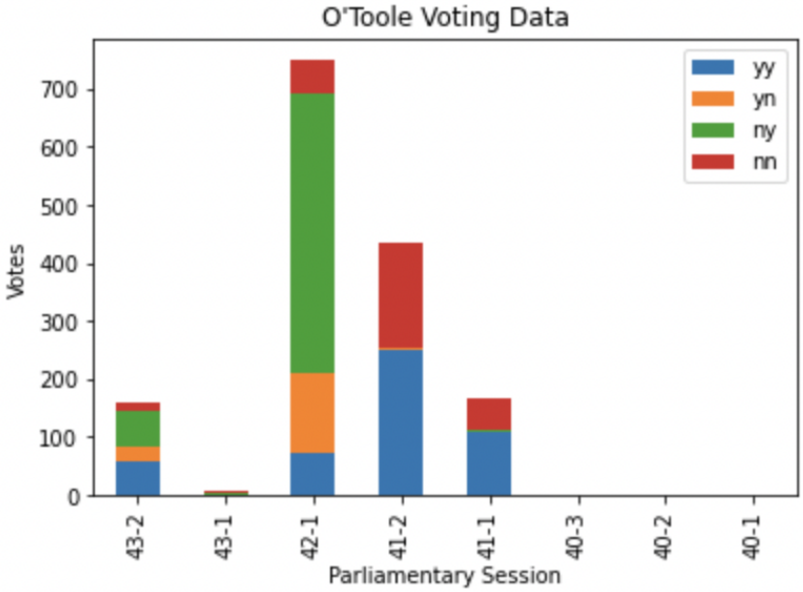


Figure 3

3. Canadian Legislators

Another plausible measurement for the performance of Canadian politicians in Parliament is to count how many bills they sponsored. To be more precise, we count the total number of sponsored bills on each topic for individual legislators. This could be an analogue of the contribution of a soccer player in a game by passing, scoring, and so on.

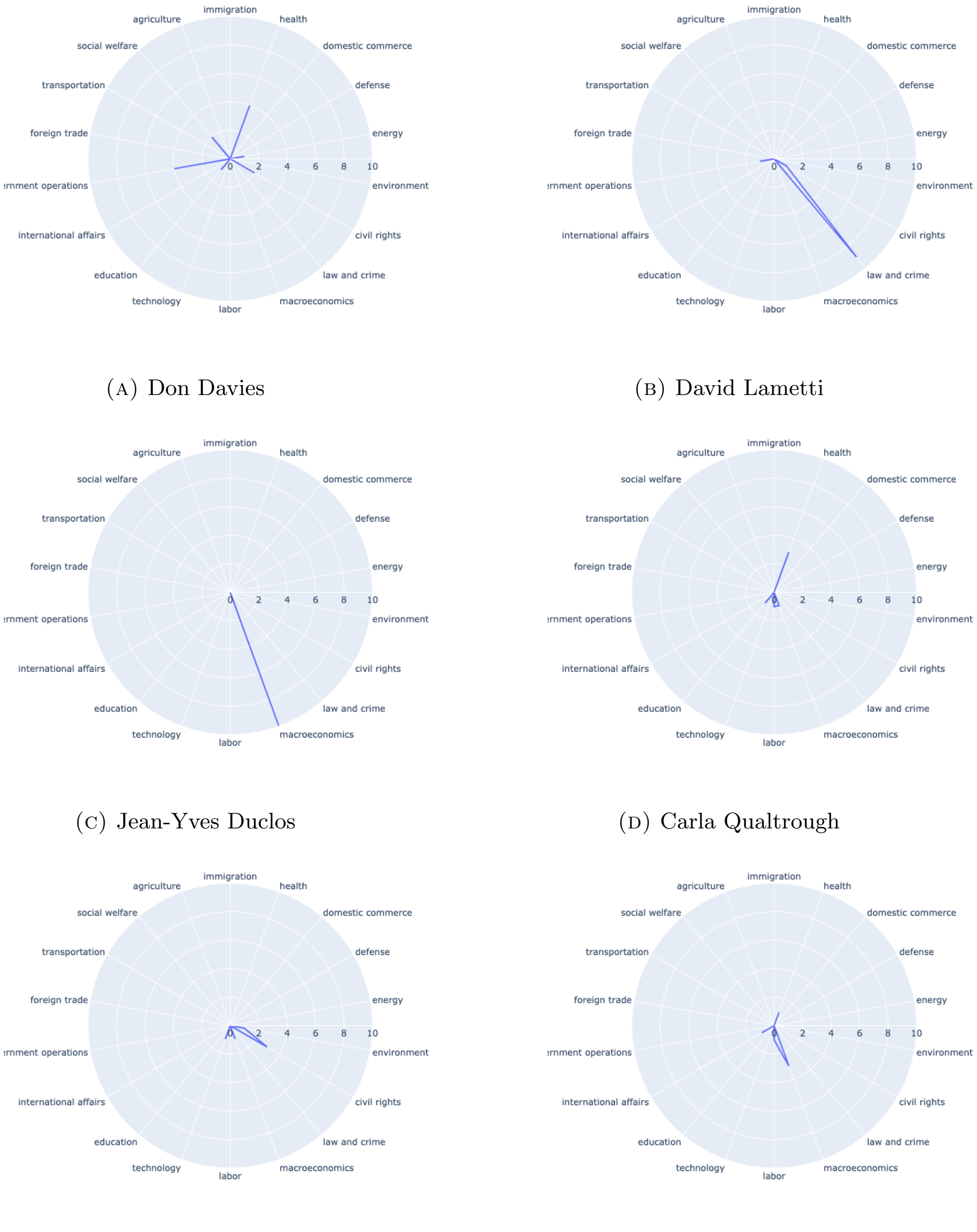
After collecting these data, we find that most legislators have completely no record. Thus this might not be a satisfying metric due to a lack of available Canadian legislation data. To visualize the performance, we draw radar plots. In Figure 4, we display the top six legislators who sponsored the most bills.

4. USA State Bill Topics

Given that the Canadian legislation data is limited, we move our focus to legislation data from the United States. But instead of considering individual legislators, this time we count the the total number of bills for each topic state by state. Due to the number of legislators might differ in different states, we use the total number of bills divided by the total number of legislators as our metric. This could be an analogue of the average performance of players in a particular position in a sport.

Doing so, from Figure 5, we see that most of states have very small graph, hence this metric is not very satisfying as well.

Nevertheless, we may still consider the weight (or percentage) for each bill topic. This measures the focus of legislation for each state, which could be an analogue of analyzing how a player in a particular position fares in a game. We use pie charts to visualize the difference between topic weights. In Figure 6, we see that



(A) Don Davies

(B) David Lametti

(C) Jean-Yves Duclos

(D) Carla Qualtrough

(E) Brian Masse

(F) Chrystia Freeland

FIGURE 4. Canadian Legislators' Radar Plot

the top two topics are always "government operations" and "law and crime", while the remaining topics differ from state to state.

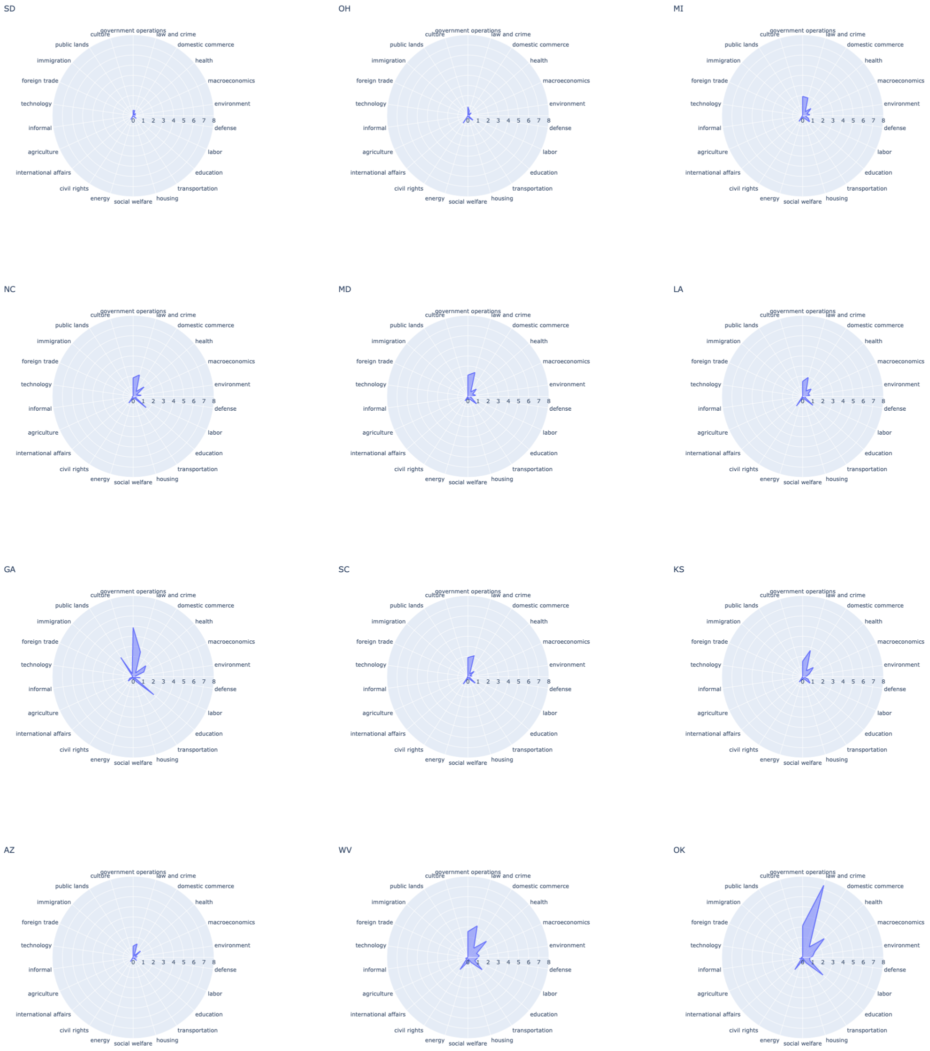


Figure 5. US State Bill Mean Radar Plot

5. Global interests versus National interests

In further pursuit of performance metrics, we turn our attention directly to issues addressed by legislators. Specifically, we focus on whether or not the issues a given legislator is concerned with is global in extent, or is confined to one's country alone.

Due to data shortage we consider two select states in the United States: Georgia and Ohio. We begin this analysis by extracting the data containing the names of legislators (state senators), the bills they sponsored, and the topics of the bills. Next, to achieve an experimental partition of the topics into those of global concern and those of national concern, we compare the topics with global issues currently

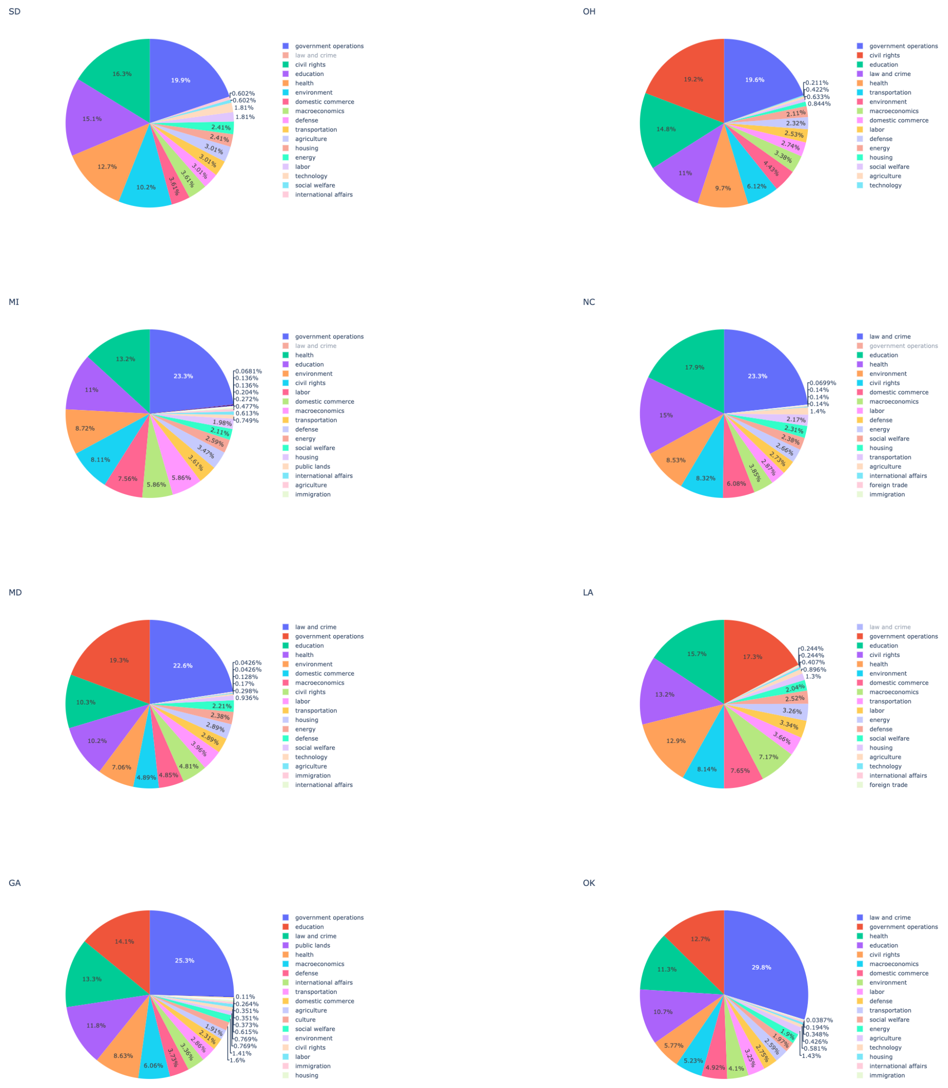
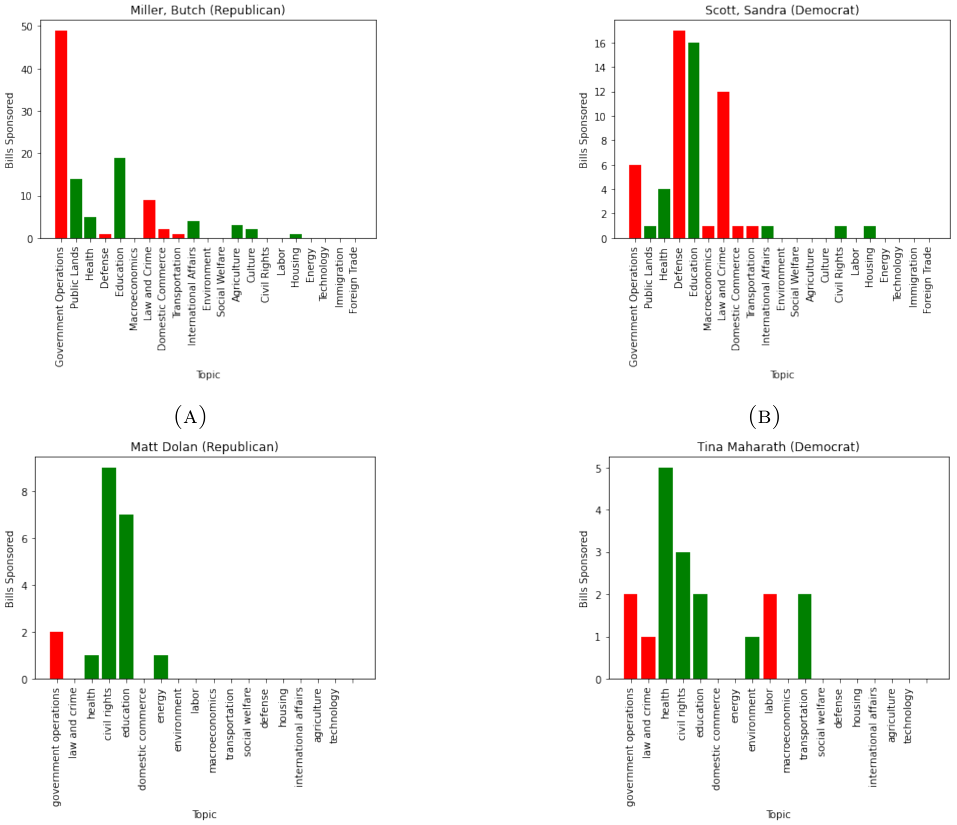


Figure 6. US State Bill weight Pie Plot

of interest to the United Nations. For example, topics such as "government operations", "defense", and "law and crime" are labelled as national issues, while "environment", "education", and "health" are labelled as global issues.

In Figure 7 we have four bar charts, each associated to a state senator (Butch Miller and Sandra Scott from Georgia, and Matt Dolan and Tina Maharath from Ohio), that track the number of bills principally sponsored by that senator under a given bill topic. The bars themselves have been labelled red and green; red indicates a topic that has been associated to a global interest, and green to a global interest.



(A)

(B)

(C)

(D)

Figure 7. Global interests versus national interests in Georgia (top row) and Ohio (bottom row)

A naïve association of political party to interests' extent might associate the Republican party to national interests and the Democratic party to global interests. This notion is challenged by this data. Indeed, the amount of red and green in the bar charts does not correlate with political party in the way that one might expect, therefore a further investigation into whether or not a metric such as this will provide a more quantitative and rigourous measure of performance may be merited.

6. Bipartisanship in the United States

We can visualize the interaction of legislators from the Democratic party and the Republican party by using network graphs. Such graphs are commonly used by social media such as Facebook, to analyze the relationships between its users. We can model each legislator as a node in a graph. Two nodes are connected by an edge if they cosponsored a bill together. In Figure 8 we look at the network graph for North Carolina legislators.

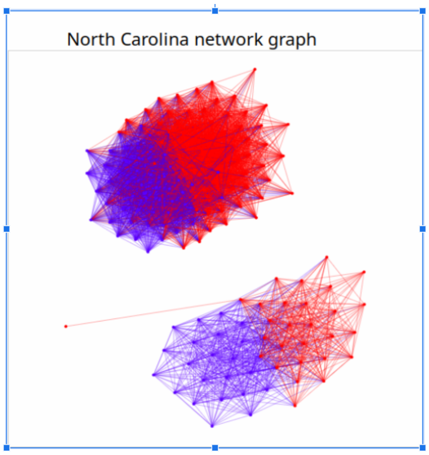


Figure 8. North Carolina Network Graph

The colour blue represents the Democratic party and the colour red represents the Republican party. We see two clusters of nodes, one where the connectivity is high within the clusters, and one where it is low across the clusters. On average, each node connects to 34.04 other nodes for members of the Democratic party, and each node connects to 31.07 other nodes for the Republican party. Democrats also have cross ratio of 30.7 and the Republicans have 23.1. All in all, this means that 30.7% of bills sponsored by Democrats are cosponsored with Republicans, and 23.1% of bills sponsored by Republicans are co-sponsored with Democrats. Upon analyzing the data sets from other states, a similar conclusion may be found: that Democrats have higher cross over ratio. This data suggests that Democrats tend to be more bipartisan.

7. Conclusion

It is fair to say that the content of this report is mostly descriptive, that is, an examination of and speculation upon historical data. Indeed, potential further work would be to gather more data and extend the preliminary metrics outlined above to rigourously evaluate performance, and perhaps even build predictive models that could illuminate directions in which politicians would carry their followers into the future. Needless to say, this is will be an extremely difficult task and it certainly requires more information and expertise, not least due to the immeasurable complexities of real world politics, where there are a lot of variables to affect politicians' votes and public opinions. Indeed, there are many things in the real world that are simply hard to quantify.

8. Acknowledgements

The authors would like to thank the organizers of the Math to Power Industry workshop for the incredible opportunity to train and acquire true industry experience. In addition, we thank our mentors Adam Kashlak of the University of Alberta, and William Spat and Samuel Leung of IOTO International for their helpful advice during the compilation of this project.

McMillan-McGee:
Characterizing Resistance and Inductance on a Copper Bus Bar

Carson Chambers, Pedro Sobrevilla-Moreno

Abstract

Voltage spikes occur when current is abruptly interrupted and can damage inadequately protected equipment. To install protective equipment, one must find the resistance and inductance of the object carrying the current. Using Maxwell's Equations one can relate the resistance and inductance to the intensity of the magnetic field induced by the current. The authors suggest a Helmholtz equation using discrete methods for computation to model the magnetic field intensity.

1 Introduction

The engineering firm McMillan-McGee came to *Math*^{industry} with a thermodynamics problem involving electromagnetic fields that required solving a certain non-homogenous boundary value problem involving Maxwell's equations. McMillan-McGee developed a high frequency inverter. This high frequency alternating current induces an electromagnetic field and if this current is abruptly interrupted then a voltage spike will occur that puts the equipment at risk of being damaged. To fix this it is necessary to install a suitable bus bar system that can absorb energy caused by switching transients from semiconductor devices. Hence, both the resistance and inductance of the DC bus bar that supplies the current must be characterized.

Previous work has been done on this subject by Norman McLachlan[1] using an ellipse to approximate a rectangular cross-section of a bus bar. In his work he developed formulas to find current density, power loss, and high frequency resistance. Unfortunately, McLachlan's[1] work is in Gaussian units instead of Meters, Kilogram, Seconds, Coulombs units, also known as MKSC, which is undesirable for an engineer. Finally, McLachlan[1] draws the conclusion that the surface distribution of current density is identical to that of a bar holding an electric charge. Hence, the total current flowing axially on the bus bar corresponds to the total surface charge.

In this report, models will be developed to characterize resistance and inductance along a bus bar in MKSC units. The models will be developed in a way such that it is trivial to change the situation dependent constants such as the wave number or length of the bus bar. Two different models will be developed, one in elliptical coordinates and the other in rectangular coordinates.

2 Theory

The bus bar being made of copper allows itself to be a good conductor which the magnetic field barely penetrates. As such, the problem can be reduced from three dimensions to two dimensions, only observing the surface layer. The component of the magnetic field normal to the bar's surface tends to decay exponentially. Thus, the magnetic field is roughly tangential to the bus bar and it follows that on the surface the vector magnetic potential is constant and satisfies the same conditions as the electrostatic potential. Using Maxwell's Equation's we can relate the intensity of the electric field to the resistance and inductance. Letting $\mathcal{D} = \epsilon_0 \mathcal{E}$, $\mathcal{H} = \frac{\mathcal{B}}{\mu_0}$, where $\mathcal{E}$ is the vector electric field, $\mathcal{H}$ is the vector magnetic field, ρ is the charge density scalar field, ϵ_0 and μ_0 are scalar values that vary depending on the units you are working in, and J is the scalar current density field, then for Maxwell's equations we have[3],

$$\begin{aligned} \frac{\partial \mathcal{D}}{\partial t} &= \nabla \times \mathcal{H} - J, && \text{(Ampère's Law)} \\ \frac{\partial \mathcal{B}}{\partial t} &= -\nabla \times \mathcal{E}, && \text{(Faraday's Law)} \\ \nabla \cdot \mathcal{D} &= \rho, && \text{(Gauss' Law)} \\ \nabla \cdot \mathcal{H} &= 0. && \text{(Coulomb's Law)} \end{aligned} \tag{1}$$

Given that we only need to concern ourselves with the surface layer of the bus bar we can look at Maxwell's Equations in two dimensions. Let $(0, 0, \mathcal{E})$ and $({}_1\mathcal{H}, {}_2\mathcal{H}, 0)$ denote the electric field and the magnetic fields, respectively. Then,

$$\mathcal{E}_y = -{}_1\mathcal{H}_t, \tag{2}$$
$$\mathcal{E}_x = {}_2\mathcal{H}_t, \tag{3}$$
$$\mathcal{E}_t = {}_2\mathcal{H}_x - {}_1\mathcal{H}_y. \tag{4}$$

In this scenario, our electromagnetic field is time harmonic. Letting i represent the complex number $\sqrt{-1}$ and k the wave number, then we have,

$${}_1\mathcal{H}(x, y, t) = e^{ikt}{}_1h(x, y), \tag{5}$$
$${}_2\mathcal{H}(x, y, t) = e^{ikt}{}_2h(x, y), \tag{6}$$
$$\mathcal{E}(x, y, t) = e^{ikt}u(x, y). \tag{7}$$

After putting these two relationship mappings together, we see by equations (2), (5) and (7) that we have,

$$e^{ikt}u_y(x,y) = -ike^{ikt}{}_1h(x,y).$$

Thus ,

$$\frac{i}{k}u_y(x,y) = {}_1h(x,y).$$

Similarly by equations (3), (6) and (7),

$$e^{ikt}u_x(x,y) = ike^{ikt}{}_2h(x,y)$$

this implies,

$$\frac{-i}{k}u_x(x,y) = {}_2h(x,y)$$

Finally, by combining the previous results and equation (4),

$$ike^{ikt}u(x,y) = (\frac{-i}{k}u_{xx}(x,y) - \frac{i}{k}u_{yy}(x,y))e^{ikt}$$

give us,

$$-\Delta u(x,y) - k^2u(x,y) = 0.$$

Notice that from Maxwells equations we have recovered a two-dimensional Helmholtz equation. Given the propensity of the electrons to cluster at the endpoints of the bus bar, we will require non-homogenous Dirichlet boundary conditions. The k value, referred to as the wave number, is determined by ω, μ and σ the circular frequency, permeability in a vacuum, and electrical conductivity of copper, respectively, with relationship $k^2 = -i\omega\mu\sigma$.

3 Methodology

The computations for this project were all done on 'MATLAB R2019a'. The boundary conditions were selected in a way to match the propensity of electrons to cluster at the endpoints of the copper bus bar. When observing the rectangular case, $(x,y) \in [0,a] \times [0,b]$, the Dirichlet boundary conditions imposed were,

$$u(0,y) = u(a,y) = \text{Amplitude} \times cos^2(\frac{\pi x}{a}),$$

$$u(x,0) = u(x,b) = \text{ Amplitude} \times cos^2(\frac{\pi y}{b}).$$

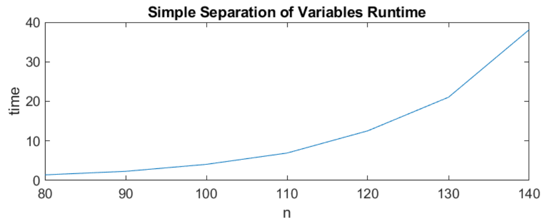
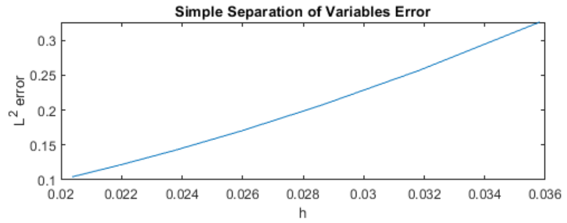
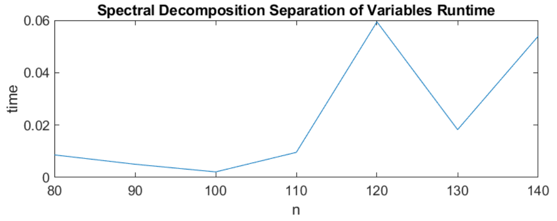
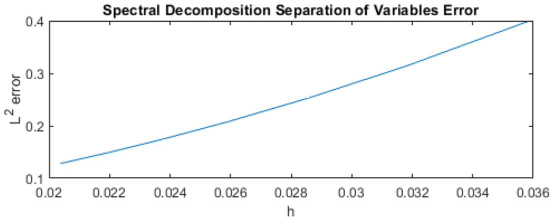
In the case of the ellipse we determined that its area should be equal to the area of the rectangle, to match the results of the rectangular case. Now, recall that an ellipse can be characterized as the points $(x,y) \ni \frac{x^2}{\gamma^2} + \frac{y^2}{\delta^2} = 1$. Hence, the boundary condition can be parameterized as $x(\theta) = \gamma cos(\theta)$ and $y(\theta) = \delta sin(\theta)$ where $\theta \in [0,2\pi)$. The initial condition can then be given by,

$u(x(\theta),y(\theta)) = Amplitude * cos^2(\theta)$.

The code has been written in such a way that it is generalized for an arbitrary problem of a similar nature. For demonstration purposes, McMillan-McGee provided us with sample values to use that matched their particular problem. The constants used are as follows (where A is amplitude),

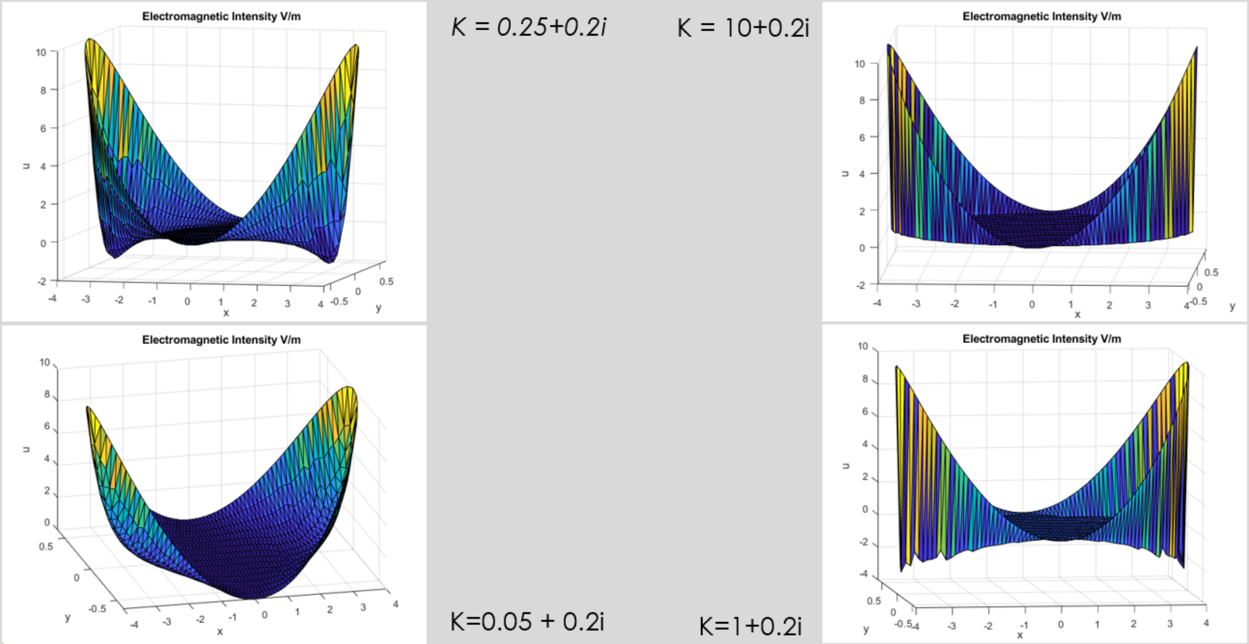
$$
\begin{aligned}
a &= 8m, \\
b &= 1m, \\
\gamma &= 4m \\
\delta &= \frac{2}{\pi}m, \\
\mu &= 4 * \pi 10^{-7}\frac{N}{A^2}, \\
\sigma &= 58.7 \cdot 10^6\frac{S}{m}, \\
\omega &= 2\pi 10^6 s^{-1}, \\
\text{A} &= 10m.
\end{aligned}
$$

With those conditions and constants, we apply the usual second-order finite difference scheme to discretize the Helmholtz equation on the rectangular domain. We solve the resultant linear system using discrete separation of variables. This method has a time complexity of $O(n^3)$ as opposed to Gaussian Elimination which has a cumbersome $O(n^6)$ time complexity[2], where $n = \frac{1}{h}$ and h is the size of the grid in each direction. In the elliptic case we applied continuous piecewise linear finite elements to solve the problem. To ensure the integrity of the data from our algorithm we tested it on an exact solution where the electric intensity was $u(x, y) = x^2 - y^2$ the wave number $k = 1$ and the forcing term $F = y^2 - x^2$. Testing both the Gaussian Elimination and discrete separation of variables we tracked both the runtime and error of each algorithm,

Runtime	Error
	
	

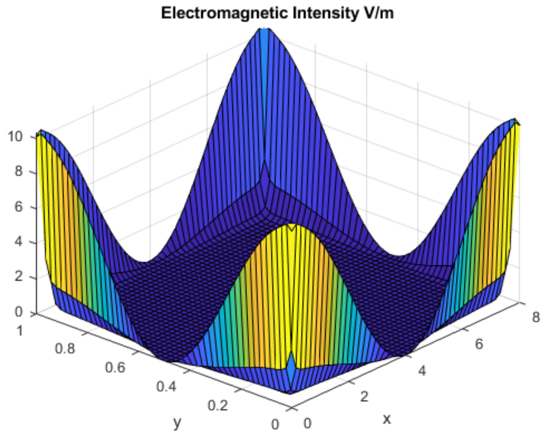
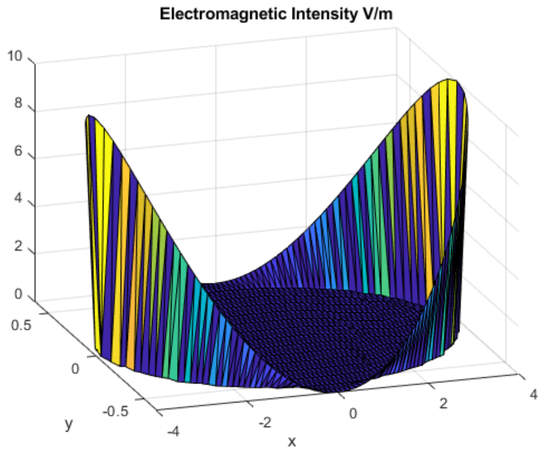
From the results we decided to go with the discrete separation of variables approach.

Additionally, we did some test runs for different values of the wave number k, to study the behaviour of our model.



4 Results

Using the equations and parameters described in the previous sections, we arrived at the following results for our two regions of interest i.e. a rectangular bus bar and an elliptical one. Firstly, we concluded that the electromagnetic intensity is greater at the end points, and decays in the midsections of the copper bus bar as shown in the figures below. This implied that our models adequately represent a real world scenario of the problem according to our discussion with our industry mentor at McGillan-McGee. Additionally, we would like to mention that from the two chosen regions used for our models, the rectangular bus bar is more applicable as it is most resembles the real world situation; since this is most common shape of bus bars. The elliptical model results will allow our industry partner at McGillan-McGee to continue investigating the work done by McLachlan[1] on the subject, which was the original problem that was presented to us. It is important to remark that our models are generalized so it is trivial to substitute the parameters for alternative conditions.

Rectangular	Elliptical
	

5 Conclusions

Having created the rectangular model that accurately depicts the scenario of current flowing through a copper bus bar, Maxwells Equations can be used to characterize resistance and inductance. Knowing the resistance and inductance will allow McGillan-McGee to alter the copper bus bar in such a way that voltage spikes will not run back through it and ruin the attached machinery. The elliptical model we created will allow McGillan-McGee to continue following along McLachlan's[1] work and potentially pursue a different avenue.

6 Acknowledgements

We would like to express our deep gratitude to our terrific mentor, Dr. Shaun Lui, for his time and wisdom. Shaun has done a terrific job of introducing us to the right materials that we needed in order to endeavour on our problem. We would also wish to thank our industry mentor, Edwin Reid, for the opportunity to work on this project with him and for his eagernes to share his industry expertise. Finally, we want to thank the organizers of the PIMS $Math^{Industry}$ workshop, especially to Professors Allen Herman and Kristine Bauer for their support.

7 References

[1]: Norman W. McLachlan, “Theory and Applications of Mathieu Functions”, Oxford University Press, 1947.

[2]: S. H. Lui, ’Numerical Analysis of Partial Differential Equations’, Wiley, 2011

[3]: J.E. Marsden, A. Tromba, "Vector Calculus Sixth Edition", W.H. Freeman and Company, 2012.

Math Industry Workshop

Natural Resources Canada   Ressources naturelles Canada

MODELLING MOUNTAIN PINE BEETLE INVASIONS BY LONG DISTANCE DISPERSAL

NOAH BOLOHAN, SAJAD FATHI HAFSHEJANI, SANTANIL JANA, AND BRYAN KETTLE

Industry mentor: Devin Goodsman, Entomologist, Northern Forestry Centre, Natural Resources Canada

Academic mentor:
Julien Arino, Professor, Department of Mathematics, University of Manitoba

Abstract. Dispersal models in mathematical biology typically represent dispersal using spatially invariant diffusion terms in partial differential equations or using dispersal kernels in integrodifference equations. In reality, dispersal is highly variable in space and time–this is especially true of long-distance dispersal. When insects disperse over long distances, they are subject to numerous meteorological variables that are dynamic in space and time. In this work, we address this challenge by abandoning the typical approaches to modelling dispersal mentioned above and instead we incorporate spatial variability using an atmospheric dispersion model driven by meteorological data. We overlay a mathematical model of beetle flight propensity and behaviour and simulate dynamics in response to changing temperature along flight trajectories driven by wind.

1. Introduction

Insect dispersal is often divided into two classes: local and long distance. Local dispersal is the most common dispersal mode. Local dispersal is well represented using dispersal kernels [6] that describe the population-level relocation patterns. Long distance dispersal is much more challenging to model because only a small proportion of insects are thought to disperse this way, and because the factors that govern whether and how far individuals disperse are meteorological [4], and therefore difficult to predict.

For example, most mountain pine beetles disperse between five and fifty meters from where they were born but approximately 0.2% traverse above tree canopies [7] where they can be pulled upwards by updrafts and then transported laterally by higher wind speeds in the lower atmosphere [4]. Long-distance dispersal is likely the dominant determinant of the speed of mountain pine beetle invasions, so describing this behaviour is of significant interest [5].

The goal of this project is pair an atmospheric dispersion model that represents the movement of air currents in three dimensional space with a stochastic differential equation model that will represent how long flying beetles remain within a

Received by the editors September 2021.

given air current. When paired, these two modelling approaches will enable realistic simulation of long-distance dispersal of the mountain pine beetle and visualization of dispersal patterns

2. Atmospheric Dispersion Model

To understand the role that meteorology plays in the long-distance dispersal of the mountain pine beetle, we used the Hysplit atmospheric dispersion model [8] developed by the American National Oceanic and Atmospheric Administration (NOAA). The Hysplit model is driven by meteorological data inputs and combines Lagrangian and Eulerian approaches to model dispersion of particles or chemicals in the atmosphere; this is the origin of the first two letters in the model name, which refer to a hybrid approach.

We capitalized on the Lagrangian aspect of Hysplit to understand how the air currents that carry mountain pine beetles may move through space and time. Trajectories of hypothetical infinitesimal volumes of air with uniform properties that are subject to meteorological conditions were modeled in Hysplit. In order to further analyze these trajectories, we utilized the Python module PySplit [9], developed by Mellissa Cross of the University of Minnesota. Pysplit introduces the HySplit model into the Python framework in order to make use of the flexible tools Python has to offer, such as analysis and plotting of trajectories.

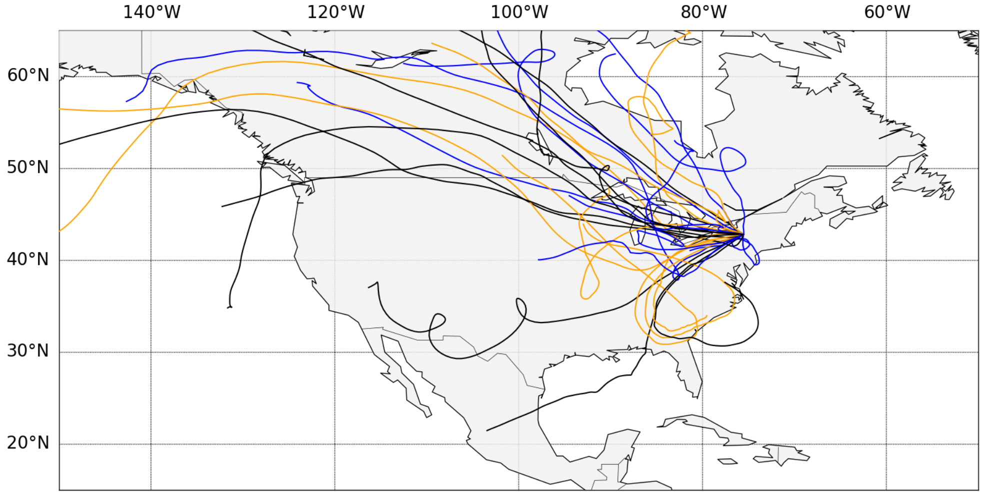


Figure 1. Example trajectories produced with PySplit, occuring during January 2007.

In Pysplit, most commonly, an initial Python script is executed in order to compute trajectories which occurred at a specific date, time, location, altitude, and with specified duration of flight. One can also choose to create various evenly spaced trajectories throughout surrounding days in Pysplit. This granted us flexibility in choosing desired meteorological conditions for air parcel trajectories. Once this step was complete, a second Python script was executed to generate trajectory plots.

For the sake of this project, we initially modeled a beetle in flight as an air parcel. Although this does not yet take into account beetle behaviour in the air, we can still obtain approximate trajectories for long-distance dispersal by estimating geographical location of beetle emergence as well as the time at which emergence likely occurred. Empirical data suggests that Mountain Pine Beetles emerged in late July and early August in 2005 [4]; in 2005, maximal emergence densities were observed on July 17, July 22 and July 26 [4]. During these periods of emergence, beetles were observed up to 800m above the forest canopy [4], where they were subject to higher wind speeds, and where they maintained flight for a maximum of 8 hours. To create the beetle trajectories that we used in combination with the stochastic differential equation model of beetle flight behaviour, we generated trajectories which occurred on July 17th, 22nd and 26th, starting at 12pm PT and progressing for 8 hours, with initial coordinates $(55.941, -121.405)$. We also considered trajectories with initial altitudes corresponding to the range in which the majority of airborne beetles were observed. These trajectories are included in Figure 2.

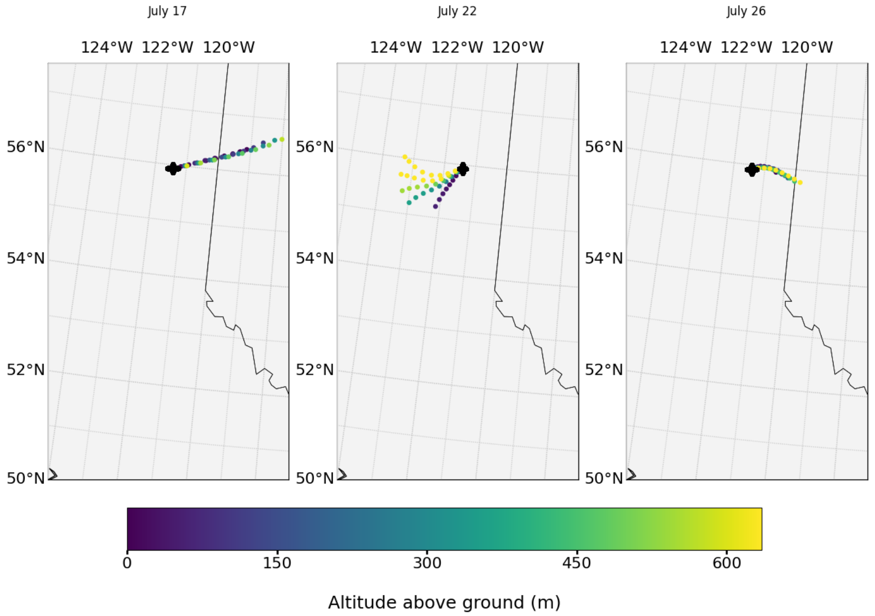


Figure 2. Approximate beetle trajectories in 2005, occurring on July 17th (left panel), July 22nd (middle panel) and July 26th (right panel). Trajectories are coloured according to altitude above ground (m), indicated by the colour bar. All trajectories begin at the black star at 12pm PT, and progress for 8 hours. Included are trajectories which have initial altitude 10m, 170m, 330m, 490m and 650m.

Although Hysplit has been used previously to model the trajectories of mountain pine beetles dispersing over long distances [1], trajectories simulated using Hysplit alone have limited utility because mountain pine beetles are not inert particles. Rather, beetles need to fly to maintain aloft: When fat reserves are depleted by the effort of flight, beetles will cease flying [2]. In addition, there are numerous behavioural responses to environmental conditions that likely result in movement patterns that differ significantly from those of inert particles. Thus, modelling long-distance dispersal of mountain pine beetles requires the addition of a level of realism by supplementing the Hysplit model simulation with a second stochastic differential equation model that represents beetle behaviour. This model is described in the following section.

3. Mathematical Model of Beetle Flight

To model the long-distance dispersal of the beetles within a moving parcel of air, we consider the mean beetle density in that parcel of air U_t as a function of time t. The mean beetle density during a flight is primarily affected by fatigue [2] and temperature. The beetles fold their wings and drop when they get tired. This process has been modeled previously as an exponential loss process such that the density of beetles decays over time like a negative exponential function [3]. Mountain pine beetles also require ideal flying conditions. When the temperature drops below their ideal flying temperature they also tend to fold their wings and drop [4]. As the process of beetle dispersal is very close to a drift-diffusion process, we use a Stochastic Differential Equation to model the mean beetle density. Considering all the factors affecting the mean beetle density U_t, we arrive at the following SDE for our model:

$$dU_t = \underbrace{-aU_t dt}_{\text{settling: fatigue}} - \underbrace{b_t U_t dt}_{\text{settling: temperature}} + \underbrace{\sigma_t U_t dW_t}_{\text{volatility}} \tag{3.1}$$

where a and b_t are the settling parameters for fatigue and temperature, σ is the percentage volatility, and W_t is a Wiener process or Brownian motion defined as follows:

Definition 1. *A standard (one-dimensional) Wiener process (also called Brownian motion) is a stochastic process $\{W_t\}_{t\geq 0}$ with the following properties:*

(1) $W_0 = 0$.
(2) *With probability* 1*, the function* $t \mapsto W_t$ *is continuous in* t.
(3) *The process* $\{W_t\}_{t\geq 0}$ *has stationary, independent increments.*
(4) *The increment* $W_{t+s} - W_s$ *has the* $NORMAL(0, t)$ *distribution.*

The fatigue parameter b_t due to temperature is dependent on time as temperatures along a trajectory changes over time.

We do not have enough beetle flight data to estimate b_t numerically. We instead try to write a function that roughly describes the settling due to temperature. First we look at the temperatures at which beetles fold their wings and which temperature is ideal for their flight at which settling rate will be very close to 0. Based on some recent research, the ideal flying temperature for beetles is more or less between $10°C$ and $35°C$. Based on these observations we take the map b_t to be

$$b_t = \begin{cases} c/e & \text{if } T_t \leq 10 \\ ce^{-\frac{1}{1-S_t^2}} & \text{if } 10 < T_t < 35 \\ 0 & \text{if } T_t \geq 35 \end{cases} \tag{3.2}$$

where $S_t = \frac{T_t - 10}{25}$ and T_t is the temperature at time t. The constant c is to be determined. The ideal scenario would be to have some flight data and be able to estimate c. For the purpose of this project we assume c to be 1. The plot of the function b_t with respect to the temperature T_t is shown in Figure 3. Notice that we have assumed that beetles don't fold their wings at temperatures above $35°C$. In the context of atmospheric dispersal, it might be a fair assumption because the warmest temperatures are going to be at ground level and beetles can disperse at temperatures of $35°C$. The parameters a and b_t are indicating deterministic trends and the parameter σ is responsible for a set of unpredictable events occurring during the beetle flight.

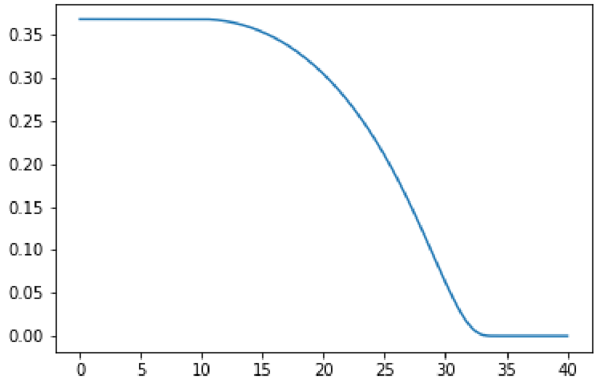


FIGURE 3. Plot of b_t (/hr) with respect to T_t ($°C$)

4. Estimation of Parameters

This section calculates the best value of parameter a representing settling rate due to fatigue. For this purpose, we first assume that beetles settle at a constant rate a. To prove this fact, we use a dataset that contains information about beetles that were flown on flight mills 2 days after they emerged and the total duration and velocity of their flight were measured [2]. To calculate the value of settling rates, we are interested to apply a negative exponential function as:

$$f(t) = \exp(-a * t), \tag{4.1}$$

where t denotes the time. Thus, a simple least squares fit allows us to obtain an estimation of the slope, that is, a. In fact, we are interested minimizing the following equation to obtain the best value for parameter a:

$$\min_a Z = \sum_{i=1}^{n} (e^{(-at_i)} - y_i)^2 \tag{4.2}$$

By taking natural log function, we have:

$$-at_i = \log(e^{(-at_i)}).$$

So, the equation (4.3) can be rewritten as:

$$\min_a Z = \sum_{i=1}^{n} (-at_i - \log(y_i))^2 \tag{4.3}$$

As we know that, the best value for parameter a can be obtained by taking the first derivative of function Z. So we have:

$$Z' = 0 \Rightarrow a = \frac{\sum_{i=1}^{n} t_i \log(t_i)}{\sum_{i=1}^{n} t_i^2} \tag{4.4}$$

Considering (4.4), we can compute settling rate due to fatigue. We used this strategy and obtained the value of parameter a for the dataset and plotted results in Figure 4.

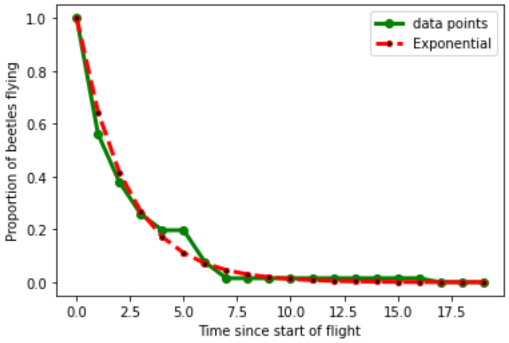


FIGURE 4. The behavior of fitting negative exponential curves

5. SOLUTION OF THE SDE

We want to find a solution of the Stochastic Differential Equation 3.1 which describes our model for the mean beetle density along a trajectory. The key step towards our solution requires an good understanding of the Itô's lemma. Let us start by stating Itô's lemma.

Lemma 1 (Itô). *For an Itô's drift-diffusion process*

$$dX_t = \mu_t dt + \sigma_t dW_t$$

and any twice differentiable scalar function $f(t, x)$ of two real variables t and x, one has

$$df(t, X_t) = \Big(\frac{\partial f}{\partial t} + \mu_t \frac{\partial f}{\partial X} + \frac{\sigma_t^2}{2}\frac{\partial^2 f}{\partial X^2}\Big)dt + \sigma_t \frac{\partial f}{\partial X} dW_t$$

The SDE we want to solve is very similar to a Geometric Brownian motion. So, we take a similar approach and define

$$Y(t) := \phi(t, U) = \ln(U_t).$$

Then by Itô's lemma

$$dY_t = \Big(\frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial U}(-a - b_t)U_t + \frac{1}{2}\frac{\partial^2 \phi}{\partial U^2}\sigma^2 U_t^2\Big)dt + \Big(\frac{\partial \phi}{\partial U}\sigma U_t\Big)dW_t \tag{5.1}$$

which simplifies to

$$dY_t = -(a + b_t + \frac{1}{2}\sigma^2)dt + \sigma dW_t$$

As the right hand side of (5) has no Y_t term, we can compute the stochastic integral:

$$\begin{aligned} Y_t &= Y_0 - \int_0^t (a + \frac{1}{2}\sigma^2)ds - \int_0^t b_s ds + \int_0^t \sigma dW_t \\ &= Y_0 - (a + \frac{1}{2}\sigma^2)t - \int_0^t b_s ds + \sigma W_s \end{aligned}$$

Substituting $Y_t = \ln(U_t)$ we have

$$\ln(U_t) = \ln(U_0) - (a + \frac{1}{2}\sigma^2)t - \int_0^t b_s ds + \sigma W_t$$

$$U_t = U_0 \cdot \exp\Big(-(a + \frac{1}{2}\sigma^2)t - \int_0^t b_s ds + \sigma W_t\Big). \tag{5.2}$$

5.2 describes the mean beetle density over time along a flight trajectory. Although, we have the integral $\int_0^t b_s ds$ in the solution, for our simulation purposes the integral will be interpreted as a Riemann sum as the temperature data along a Hysplit trajectory is discrete, not continuous.

6. Results and Discussion

We simulate the beetle flights along some of the Hysplit trajectories from section 2 using the solution 5.2 of the SDE. We have different Hysplit trajectories on different heights above the ground level. We will run our simulations along the Hysplit trajectories with height 330 meters and 170 meters from ground level. We also have Hysplit data for three different dates on which the average temperature along the beetle flight are different.

Based on preliminary examination of the beetle flight data and prior analyses, a good estimate of the parameter σ is 0.68/hr. In section 4 from our estimates for the parameter a, we see that the ideal value of a lies between 0.2/hr and 0.3/hr. For our first simulation we choose a to be 0.24/hr and run the simulation along the three Hysplit trajectories at 330 meters above the ground level on three different dates. The plots of the mean beetle density with respect to time for our first simulation is described is Figure 5. The key observation from this simulation is that the number of beetles drop faster when the mean air temperature is lower.

We assumed the initial beetle number in an arbitrary parcel of air to be 100. This number is somewhat random. As we are dealing with beetle density we can normalize the y-values to be between 0 and 1. In that way the y-axis can be interpreted as beetle density.

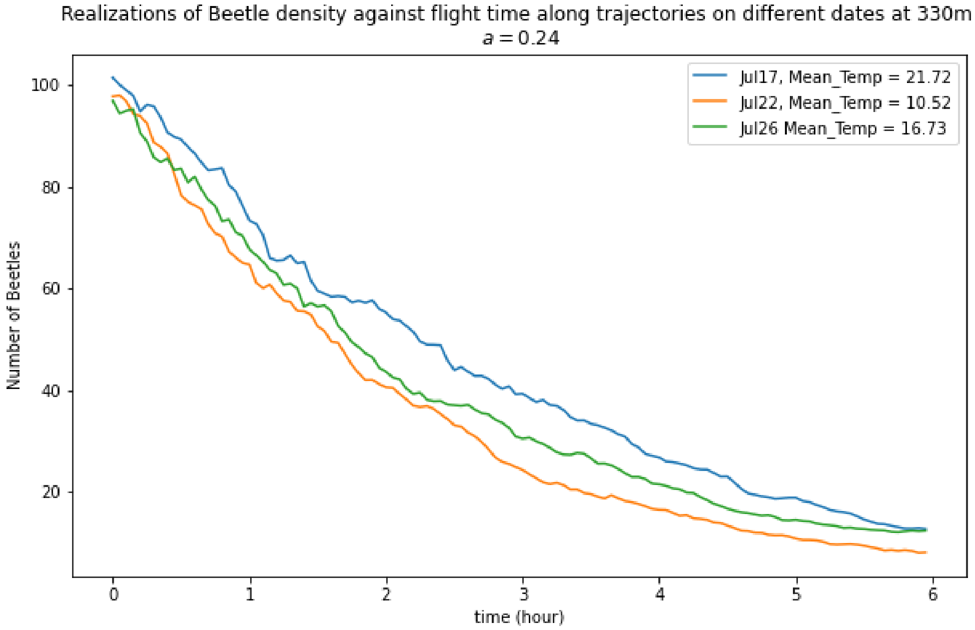


Figure 5. Beetle numbers along trajectories on three different dates

For our second simulation we take three different values for a, namely 0.2/hr, 0.24/hr, 0.28/hr and run the simulation along the Hysplit trajectory on Jul 17 at 170 meters above the sea level. We see that the number of beetles drop faster when the value of the settling parameter a is higher, which is expected (see Figure 6).

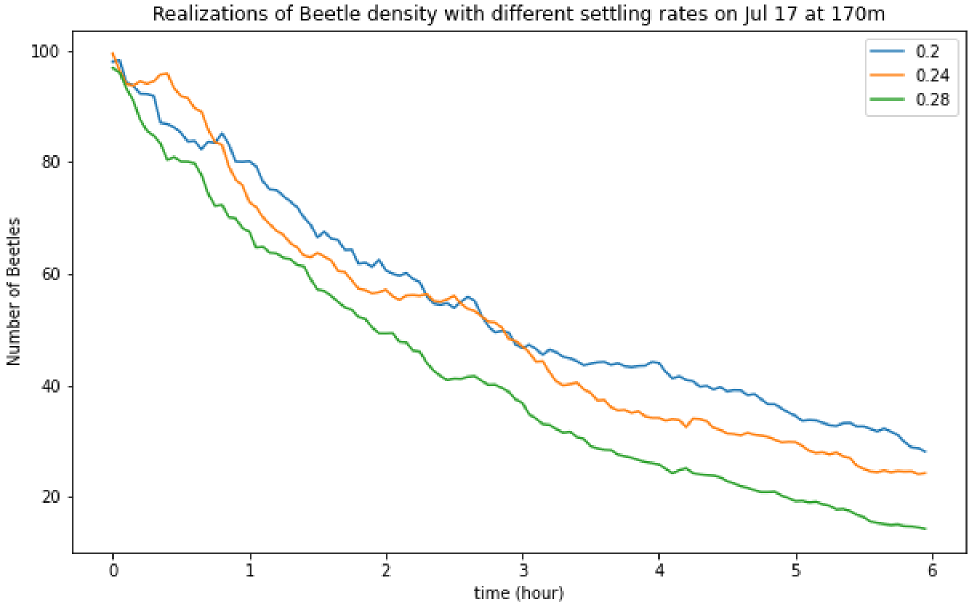


Figure 6. Beetle numbers along a trajectory with three different a-values

Acknowledgements

The team would like to thank our industry mentor Devin Goodsman and our academic mentor Julien Arino for their guidance, support, and feedback during the advancement of this project. We are also incredibly thankful to Allen Herman, Kristine Bauer, Ian Allison, and the rest of the PIMS organizing committee and instructors of the M2PI 2021 workshop for providing us this great experience.

References

[1] Bruce Ainslie and Peter L Jackson. "Investigation into mountain pine beetle above-canopy dispersion using weather radar and an atmospheric dispersion model". In: *Aerobiologia* 27.1 (2011), pp. 51–65.

[2] Maya L Evenden, CM Whitehouse, and J Sykes. "Factors influencing flight capacity of the mountain pine beetle (Coleoptera: Curculionidae: Scolytinae)". In: *Environmental entomology* 43.1 (2014), pp. 187–196.

[3] Devin W Goodsman et al. "Aggregation and a strong A llee effect in a cooperative outbreak insect". In: *Ecological Applications* 26.8 (2016), pp. 2623–2636.

[4] Peter L Jackson et al. "Radar observation and aerial capture of mountain pine beetle, Dendroctonus ponderosae Hopk.(Coleoptera: Scolytidae) in flight above the forest canopy". In: *Canadian Journal of Forest Research* 38.8 (2008), pp. 2313–2327.

[5] Kelsey L Jones et al. "Factors influencing dispersal by flight in bark beetles (Coleoptera: Curculionidae: Scolytinae): from genes to landscapes". In: *Canadian Journal of Forest Research* 49.9 (2019), pp. 1024–1041.

[6] Mark Kot, Mark A Lewis, and Pauline van den Driessche. "Dispersal data and the spread of invading organisms". In: *Ecology* 77.7 (1996), pp. 2027–2042.

[7] L Safranyik et al. "Dispersal of released mountain pine beetles under the canopy of a mature lodgepole pine stand". In: *Journal of Applied Entomology* 113.1-5 (1992), pp. 441–450.

[8] AF Stein et al. "NOAA's HYSPLIT atmospheric transport and dispersion modeling system". In: *Bulletin of the American Meteorological Society* 96.12 (2015), pp. 2059–2077.

[9] Mellissa SC Warner. "Introduction to PySPLIT: A Python toolkit for NOAA ARL's HYSPLIT model". In: *Computing in Science & Engineering* 20.5 (2018), pp. 47–62.

University of Ottawa
E-mail address: nbolo094@uottawa.ca

University of Lethbridge
E-mail address: sajad.fathihafshejan@uleth.ca

University of British Columbia
E-mail address: santanil@math.ubc.ca

University of Alberta
E-mail address: bkettle@ualberta.ca

Math Industry Workshop

Natural Resources Canada   Ressources naturelles Canada

OPTIMIZING MARKING TECHNIQUES FOR MARK-RECAPTURE STUDIES OF MOUNTAIN PINE BEETLES

J. BENESH, D. GOODSMAN, R. HAN, J. HOEPNER, H. HUANG, AND M. RAY

Industry mentor: Devin Goodsman, Entomologist, Northern Forestry Centre, Natural Resources Canada

Academic mentor:
Hui Huang, PIMS Postdoctoral Fellow, University of Calgary

Abstract. Whereas mark-recapture studies are sometimes used to estimate population size, mark-recapture studies initiated by Natural Resources Canada attempt to estimate movement of individuals in the population by marking them at source locations and recapturing them at various surrounding trap sites. A novel variation of traditional mark-recapture techniques developed by Natural Resources Canada involves coating trees with paper that fluoresces under black light such that the beetles are marked with paper dust as they emerge. Recaptured beetles are then photographed under black light. In this work, we classify images of the recaptured beetles as marked or unmarked using deep neural networks. In particular, we use transfer learning where existing top-performing classifiers are applied to our beetle image classification problem. We compare the performance of ResNet50 and EfficentNet base models when applied to images processed in a variety of ways.

1. Introduction

Since 1990, an outbreak of the mountain pine beetle (*Dendroctonus ponderosae*) has affected over 20 million hectares of forest in western Canada, making it the largest recorded insect outbreak in North American history. Mountain pine beetle adults disperse to attack and colonize trees in order to lay eggs beneath the outer bark. The process of attack and colonization disrupts nutrient flow and results in tree death [2, 5]. Although understanding beetle dispersal in this context is vital in making well-informed environmental decisions, tracking beetle relocation when adults disperse is challenging. Natural Resources Canada (NRCan) recently initiated a study designed to develop new and improved methods to quantify mountain pine beetle dispersal and to understand how far dispersing mountain pine beetles fly.

Mark-recapture studies typically involve the application of a harmless indicator to a small number of individuals, which are then released back into

This work was funded by PIMS and Mitacs.

the general population. The likelihood of recapturing a marked individual is thus inversely proportional to the size of the population, assuming nearly all of the marked individuals are still alive, and provided no significant immigration into or out of the population has occurred between the release and recapture dates. The goal of mark recapture studies initiated by NRCan is slightly different: Marked beetles are recaptured at various locations from release sites in order to better understand the dispersal process [6].

A recently developed NRCAN marking technique involves covering trees in paper that fluoresces under black light such that the beetles are coated in paper dust as they emerge, thereby allowing the marked beetles to naturally disperse without direct human intervention. Mountain pine beetles emerging from papered trees and control trees were later captured and photographed under black light.

Since manually classifying each image as marked or unmarked can be tedious and prone to error, it would be beneficial to automate the process using machine learning. The goal of this project is to develop a new algorithm to identify marked beetles by optimizing pre-existing image classification techniques. These techniques are discussed in further detail in Section 2, and their implementation is presented in Section 3. In Section 4, the project is summarized and potential improvements to the algorithm are addressed.

2. Deep Neural Networks

Deep learning has been well justified by its tremendous empirical success and state-of-the-art performance on various relevant real-life applications such as speech recognition [3], image recognition [4], language translation [10], and as a novel method for scientific computing [1]. It is an approach that enables the realization of complex tasks such as the ones mentioned above, by means of highly parameterized functions, called deep artificial neural networks $\mathcal{N} : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$. A classical architecture is the one of feed-forward artificial neural networks of the type

$$\mathcal{N}(x) = \sigma\left(W_L^\top \sigma\left(W_{L-1}^\top \dots \sigma\left(W_1^\top x + b_1\right) \dots\right) + b_L\right) , \tag{1}$$

where L is depth of the network, the function σ is a scalar activation function acting component-wise on vectors, for each layer $\ell = 1, \dots, L$, the matrix $W_\ell \in \mathbb{R}^{d_{\ell-1} \times d_\ell}$ represents a collection of weights, and the vector b_ℓ represents shifts/biases. The neural network (1) is then trained to minimize a given loss function (e.g, Mean Squared Error, Cross-Entropy, Kullback-Leibler divergence, or Wasserstein distances) over the parameters (weights and biases) of the network, usually measuring the misfit of input-output information over a given finite number of labeled training samples.

In this report we use Convolutional Neural Networks (CNNs) to solve our image classification problem. However, to train on a very large dataset, deep CNN models may take a significant amount of time. A way to bypass this process is to re-use the model parameters from pre-trained top performing

CNN models that were developed for standard computer vision benchmark datasets, such as the ImageNet image recognition tasks. This is the so-called transfer learning. One can see it as the deep learning version of "standing on the shoulder of giants". There are many top-performing models that are available for the basis for image recognition tasks, such as VGG (e.g. VGG19 [7]), GoogLeNet (e.g. InceptionV3 [8]), Residual Network (e.g. ResNet50 [4]) and EfficientNet (e.g. EfficientNetB0 [9]). In the following we are going to focus on the implementations on the Residual Network and the EfficientNet models, which are the state-of-the-art methods in imagine classification.

3. Exploration of the dataset

Our dataset consisted of 1057 images, each of which was labeled in 5 parts.

- The first component of each name is whether or not the tree segment from whence the beetle in the image emerged was **papered or not**. Papered bolts were considered marked whereas unpapered bolts can be considered unmarked.
- The second component of each name is the **color** of the paint that was applied to the outside of the trees from which beetles emerged: Possible values of the paint color include: transparent, green, pink, or control (no paint).
- The third component of each name is whether the marked and unmarked beetles were **mixed** in a jar together to test the persistence of the marking paper in a slightly more realistic context. We note that mixing could potentially lead to cross-contamination of unpapered beetles as a result of physical contact with marked beetles or with paper fragments that may have been shed from them.
- The fourth component is a **number** that is not unique to each beetle.
- The fifth and final component indicates whether beetles were photographed on their **dorsal (d) or ventral (v) sides**. We note that the tips of the abdomens and the mandibles on the ventral side are a location of higher concentrations of paper particles in some cases when beetles were marked.

A final component that was added to some of the images were comparison images which were labeled "light". These images were removed from the dataset as they were not imaged under blacklight and had a very distinct look to them (see Figure 1) in comparison to the rest of the images. This brought our total count down to 1013 images.

Finally, a series of images beginning with the word "Trap", which indicates that the beetles originated from a separate outdoor experiment in which standing trees that were infested with mountain pine beetles were papered. A number of Lindgren funnel traps were set up in the vicinity to capture beetles emerging from trees in the area. Most of the trapped beetles

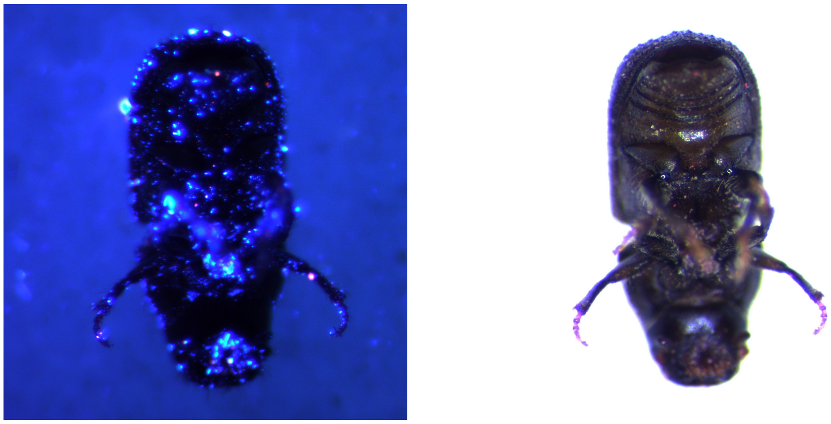


Figure 1. Blacklit vs. Lit Beetle

likely emerged from unmarked trees. For these beetle (and other insects), we do not know whether they emerged from papered trees. We chose to use this set of images as a validation set to see how well our machine learning algorithm of choice would be able to inform whether there are any marked beetles in the Trap set as well as the associated probabilities of being marked for each Trap insect potentially. From visual inspection, with the human classifier, we did not believe there were any marked beetles in this set.

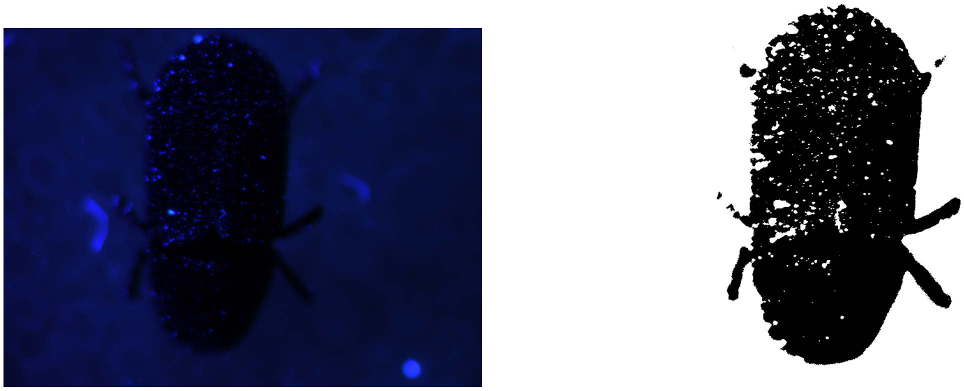
Using this labeling scheme we determined that we had 735 marked images, and 278 unmarked images, which is quite an unbalanced dataset. To mitigate this imbalance we chose to remove all images with the "Pink" tag, which fluoresced quite brightly under black light, and which formed the largest subset of our marked images. Upon removing these images, we were left with a total of 480 marked images, and 278 unmarked images, which gave us a much better balance to begin training our machine learning algorithms on.

We had two versions of this final dataset of 757 images - one with the original images and one with black and white images. The latter version was created using thresholding to convert each image into one with a binary colour scheme and then cropping around the beetle. We therefore suggestively refer to this data set as "Threshcropped". See Figure 2 to compare a sample from each data set.

4. Implementation and results

We ran our models with both the original and threshcropped datasets. We use the confusion matrix to visualize and test the performance of these models. The accuracy can be computed by the formula

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{Total}}.$$



(A) Original Image

(B) Threshcropped Image

Figure 2. Two versions of the dataset

The true positive rate is given by

$$\text{True positive rate} = \frac{\text{True positives}}{\text{Actual positives}}$$

We briefly describe each type of model below.

4.1. **Human Classifier.** A human classified the original images as marked or unmarked (see its confusion matrix in Figure 5 right). The prediction accuracy is

$$\text{Accuracy} = \frac{321 + 256}{755} \approx 76.4\%,$$

and the true positive rate is

$$\text{True positive rate} = \frac{321}{321 + 156} \approx 67.3\%.$$

Note here that there are two images missing because we have trouble loading them. We compare all other models' performances to this baseline.

4.2. **ResNet50 using original images.** *ResNet50* [4] is one of the most powerful deep neural networks which has won the ILSVRC 2015 competition because of its fabulous performance. It was proposed to solve the issue of vanishing/exploding gradient phenomenon. The idea is to use the "Residual Block" (see Figure 3) to skip connections and after-addition activations. If we denote by $\mathcal{F}(x) = \sigma(W^T x + b)$ a generic layer of the network, then the residual layer can be described as

$$x^{n+1} = x^n + \mathcal{F}(x^n). \tag{2}$$

Now let us present the transfer training result through ResNet50, which takes around 3 hours to run. First we refer to Figure 4 for the accuracy and loss evolution in terms of epochs on both training and validation set.

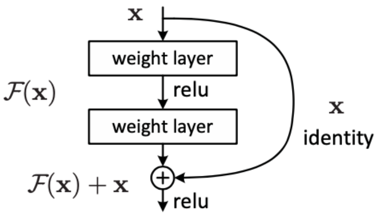

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 3. Residual Block, see [4]

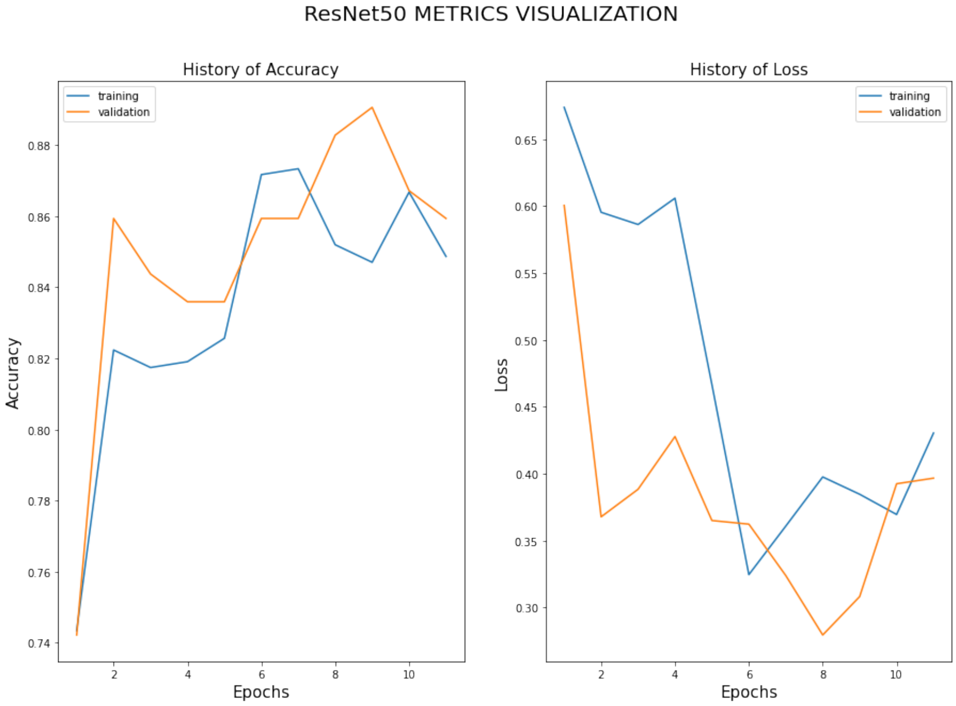


Figure 4. Accuracy and loss evolution in terms of epochs on training set (80%) and validation set (20%) from ResNet50.

It indicates that the accuracy can be up to 87.3% for the training set and 89.1% for the validation set.

Next we use the confusion matrix to test the performance of our classification model obtained from ResNet50. As it is showed in Figure 5 (left) that we have 431 true positives (marked beetles were being predicted marked), 48 false negatives (marked beetles were being predicted unmarked), 29 false positives (unmarked beetles were being predicted marked), and 249 true negatives (unmarked beetles were being predicted unmarked). The overall

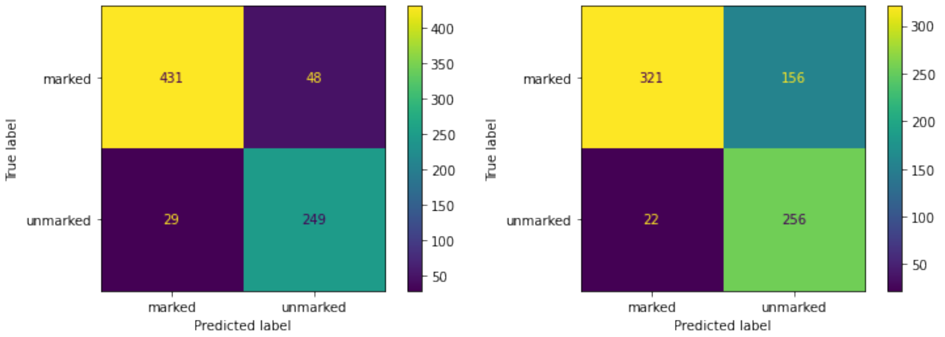


Figure 5. Left: Confusion matrix for the beetle classifier from ResNet50. Right: Confusion matrix for human classifier with the naked eye.

prediction accuracy is approximately 89.8% and the true positive rate is 90.0%.

4.3. **ResNet50 with preprocessing.** We use the sigmoid activation function at the output layer for the binary image classification. For each image, the model predicts the probability of being in the 'marked' class. Therefore, if the prediction of an image is greater than or equal to 0.5, we assign it the label 'marked'. Otherwise, we assign 'unmarked'. The validation accuracy is approximately 84.2% which is a little bit lower than the one without preprocessing, whose accuracy is 88.2%.

The confusion matrix summarizes the predictions made on the validation set in Figure 6.

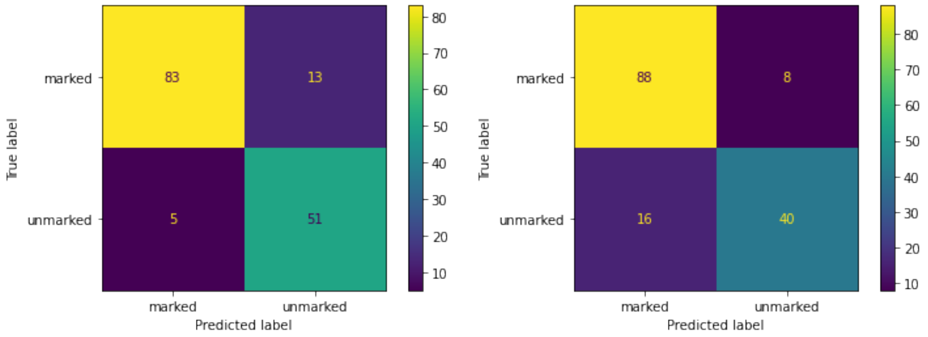


Figure 6. Confusion matrix for the validation set from ResNet50. Left: original images; Right: preprocessed

4.4. **EfficientNet.** *EfficientNet* was first introduced in [9], since then it has became one of the most efficient models that reaches state-of-the-art accuracy on both ImageNet and common image classification transfer learning

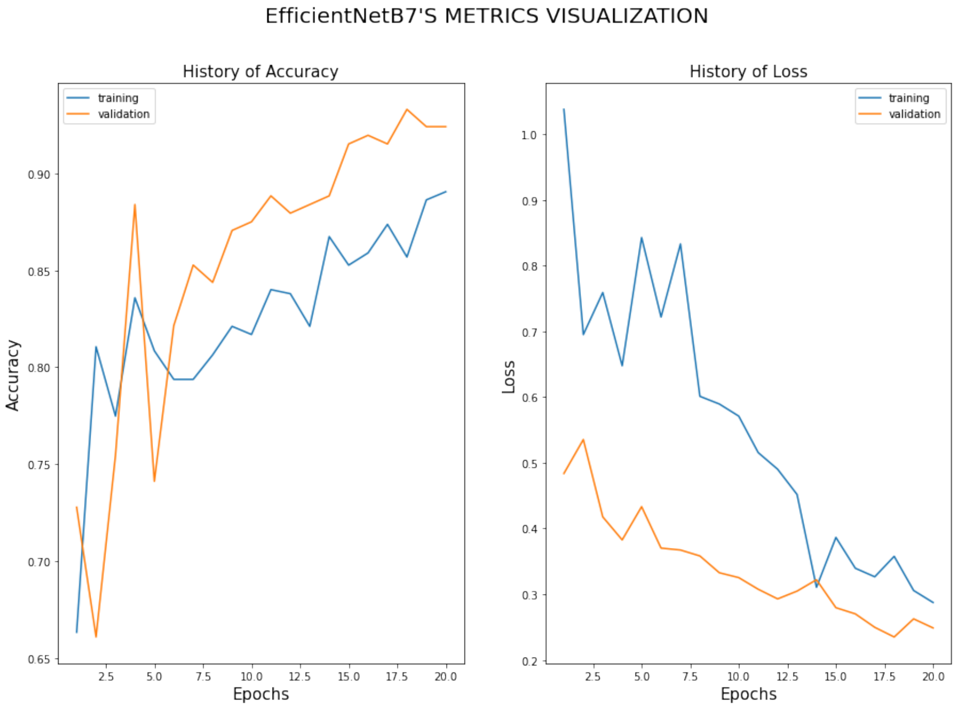


Figure 7. Accuracy and loss evolution in terms of epochs on training set (67%) and validation set (33%) from EfficientNetB7.

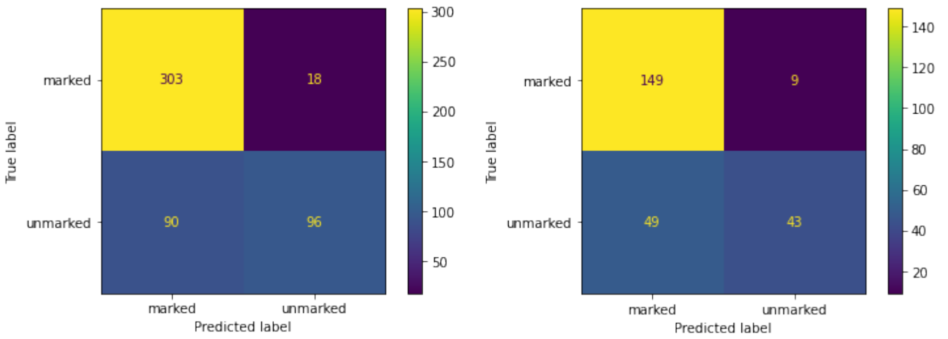


Figure 8. Left: Confusion matrix for the training set from EfficientNetB7. Right: Confusion matrix for the validation set from EfficientNetB7.

tasks. It proposes a compound scaling method to scale up CNNs in order to obtain better accuracy and efficiency. Unlike conventional approaches that arbitrarily scale network dimensions, such as width, depth and resolution, the EfficientNet uniformly scales each dimension with a fixed set of scaling coefficients. More specifically, it uses a compound coefficient φ to uniformly scales network width, depth, and resolution in a principled way:

$$\begin{aligned} \text{depth: } & d = \alpha^{\phi} \\ \text{width: } & w = \beta^{\phi} \\ \text{resolution: } & r = \gamma^{\phi} \\ \text{s.t. } & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ & \alpha \geq 1, \beta \geq 1, \gamma \geq 1\,, \end{aligned}$$

with α, β and γ to be determined by a grid search. Especially for our image classification problem, depending on the choice of the resolution of the input image, we can use a series of EfficientNet models from B0 to B7.

In the following we use the EfficientNetB7 with resolution 600 × 600 for our transfer learning, since our given images have very high resolution (1944 × 2592). Compared to ResNet50, it takes at least 6 hours to run the EfficientNetB7. From the evolution of the accuracy and loss (see in Figure 7), one can see that the accuracy can be as high as 93.3%. However this is not consistent with the confusion matrices built from the training data and validation data by using the model we have trained. Indeed, it follows from Figure 8 that the prediction accuracy of the training set is

$$\frac{303 + 96}{507} \approx 78.7\%\,,$$

and the prediction accuracy of the validation set is

$$\frac{149 + 43}{250} \approx 76.8\%\,.$$

Especially the performance on predicting unmarked beetles is very poor. Different versions of EfficientNet were run on both the original and threshcropped images. We defer the reader to Table 1 in section 5 to see the performance metrics.

5. Conclusion

The following table summarizes the results from running different models. We saw that generally, we could classify beetle images much more accurately

Model	**Dataset**	**Validation Accuracy**	**Training Accuracy**
Human Classifier	Original	76.4%	N/A
Efficient B7	Original	93.3%	89.1
Effiecient B7	Threshcrop	78.3%	85.6
Efficient B5	Original	88.0%	85.9
Resnet 50	Original	89.0%	87.3
Resnet 50	Threshcrop	84.2%	80.0

Table 1. Summary of model performance

and efficiently using machine learning tools compared to the human eye. The models ran faster with threshcropped images but with lower accuracy than the the model trained with original images. Overall, ResNet50 with original images performed the most effectively and consistently. We recommend training with preprocssed images when the computation resources are limited as the training can be faster with reasonable accuracy; otherwise, if computation resources allow and higher accuracy is desired, using the original images is suitable.

The major constraints we faced during the project were naturally those of time and resources. We were able to train three models with two datasets (original and preprocessed). More computational power such as cloud GPU should significantly improve the number of models we could train with ease and yield more information about how to make them better. Furturemore, training models with more computational power would enable us to gain insight into the optimal hyper-parameters such as the learning rate and the number of hidden layers.

Moving forward, one can try a number of things to improve predictions. This includes balancing out the training and validation datasets - removing the “Trap” and “mixed” images from the dataset and then choosing the 278 unmarked images and comparing them with a randomly chosen set of 278 beetles from the “marked” set. One can also make the training process more specialized - only train using the ventral sides. The reason for choosing the ventral side has to do with beetle behaviour and biology, which increases the likelihood of collecting paint particles on the mandibles and tips of the abdomens on the ventral side. Initial experiments in this direction looked promising. Finally, one can vary model parameters such as number of layers, image size, etc to experiment with further optimizations.

References

[1] J. Berner, P. Grohs, and A. Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. *SIAM Journal on Mathematics of Data Science*, 2(3):631–657, 2020.

[2] A. Dhar, L. Parrott, and C. D. Hawkins. Aftermath of mountain pine beetle outbreak in British Columbia: Stand dynamics, management response and ecosystem resilience. *Forests*, 7(8):171, 2016.

[3] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] L. Safranyik. Mountain pine beetle: biology overview. In *Proceedings: Symposium on the Management of Lodgepole Pine to Minimize Losses to the Mountain Pine Beetle. USDA Forest Service, Intermountain Forest and Range Experiment Station, Gen. Tech. Rep. INT-262*, pages 9–12, 1989.

[6] L. Safranyik, D. Linton, R. Silversides, and L. McMullen. Dispersal of released mountain pine beetles under the canopy of a mature lodgepole pine stand. *Journal of Applied Entomology*, 113(1-5):441–450, 1992.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[9] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Joel Benesh, Department of Mathematics and Computer Science, University of Lethbridge
E-mail address: Joel.Benesh@uleth.ca

Devin Goodsman, Entomologist, Northern Forestry Centre, Natural Resources Canada
E-mail address: goodsman@ualberta.ca
E-mail address: devin.goodsman@canada.ca

Rachel Han, Department of Mathematics, University of British Columbia
E-mail address: 13hanr@gmail.com

Jules Hoepner, Department of Mathematics, University of Victoria
E-mail address: julesihoepner@gmail.com

Hui Huang, Department of Mathematics and Statistics, University of Calgary
E-mail address: hui.huang1@ucalgary.ca

Mishty Ray, Department of Mathematics and Statistics, University of Calgary
E-mail address: mraysamar@gmail.com

Math INDUSTRY

serious:labs®

MODELING REAL-TIME HYDRAULIC SYSTEMS WITH POSITION-BASED DYNAMICS

MAKSYM NEYRA-NESTERENKO, MOUMITA SHAU, BAHAR MOUSAZADEH, AND ARNAUD NGOPNANG

Abstract. Real-time simulation of various complex physical phenomena is a challenging task, since one must meet the demands of limited computing time for high frame rates and accurate, reproducible physical simulation. One approach to balance efficient and accurate simulation is to model the dynamics with position-based dynamics. In this report, we investigate how one can adapt position-based dynamics to incorporate hydraulic pressures to simulate heavy equipment. The key impact of this work is to lay groundwork for developing standardized virtual training and evaluation of heavy equipment operation and safety.

1. Introduction

With the advent of virtual reality technologies, virtual reality simulation and equipment offer great potential in providing a high degree of realism to simulate a dynamical system. An unavoidable bottleneck for providing this degree of realism is the computing time available for real-time simulation. Visualizing the environment requires fixing a frame rate, and depending on the available computing resources, there is a limited number of calculations that can be done in one frame. When simulating dynamics, using high accuracy solvers and good choice of time step size achieves the desired precision, but at the cost of more computing time. For real-time simulation, the frame rate is known or prescribed, so we are limited by the time complexity of methods we can use to simulate dynamics. Our goal is to devise a real-time simulation framework that simulates both physical bodies (particle-based) and fluid pressures (hydraulics) with the intent of simulating heavy equipment. The prospects of this are to provide operation and safety training in virtual reality.

To develop a hybrid particle-based and fluid pressure simulation, we start from position-based dynamics (PBD) [4]. We provide insight on how PBD can be used to incorporate fluid pressures and provide a sufficiently accurate and stable simulation. The difficulty lies in accounting for the latter. Accuracy is bottlenecked by a computing time budget and stability is affected by the stiffness of a hydraulic system model. In the next sections, we proceed as follows. First, we discuss hydraulic system modeling. Second, we provide an overview of the classic PBD algorithm, which offers a simple and fast algorithm for dynamics simulation. Finally, we briefly discuss our results and future work.

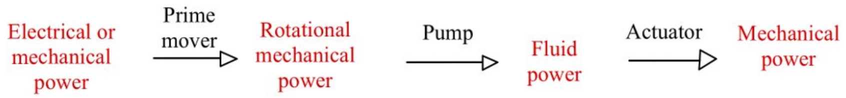


Figure 1. Flow diagram of energy transfer through a hydraulic system.

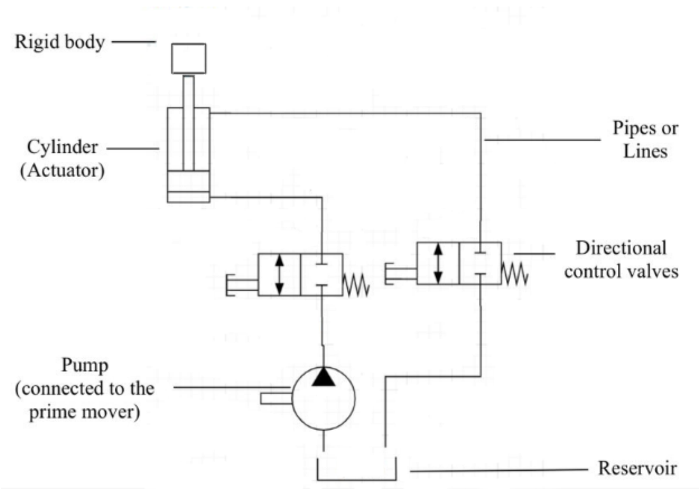

Rigid body
Cylinder (Actuator)
Pipes or Lines
Directional control valves
Pump (connected to the prime mover)
Reservoir

Figure 2. Example of a hydraulic system schematic.

2. Related work

2.1. **Hydraulic systems.** A hydraulic system aims to apply a force at a point and transfer it to another point using a liquid. This leverages the underlying physics by having large forces be produced by small input forces. For example, the structure of a hydraulic system can be described as in Figure 1. An electrical or mechanical power can be used to facilitate the movement of a pump. This produces rotational mechanical power which generates fluid flow. The flow is propagated to an actuator, which produces a force to create mechanical movement. This could be, for example, moving a hydraulic cylinder.

Hydraulic systems have numerous uses, and are abundant in the design of construction equipment, vehicles, and manufacturing equipment. In practice, hydraulic systems are described using schematics with a standardized set of notation and symbols. Figure 2 shows a hydraulic system schematic with its component symbols labelled.

The process of modeling a hydraulic system is to: first identify all important components of the system, such as pumps, valves, orifices, cylinders, motors, and so on. Next, the components are connected using a schematic, with a volume associated with each component. From there, we set the differential equations governing the system, with the unknown variables consisting of pressures, flow rates, and object positions.

2.2. Position-based dynamics.

The most popular approaches to real-time simulation of dynamical systems are based on forces. In practice, this is done as follows. Given a fixed frame or time, first identify and calculate all internal and external forces in the system. Next, compute the accelerations of all the objects using Newton's second law of motion. Lastly, perform numerical integration to compute and update the velocities and positions of the objects. However, there are alternative approaches. The one we are interested in is *position-based dynamics (PBD)* [4], which directly works with and manipulates positions. The PBD algorithm offers a robust method for simulation of deformable bodies. PBD is also very simple and computationally efficient, and thus suitable for real-time applications.

The main features of PBD are

- control over explicit integration,
- using and handling purely positional constraints,
- and direct manipulation of vertex positions.

The positional constraints are enforced using a Gauss-Seidel solver, by first predicting future positions of the points from external forces and previous velocity, and then performing a cost-effective projection of the constraints. However, there are a few notable issues. First, this framework results in loss of momentum conservation. Second, PBD's control over stiffness depends nonlinearly on the time step size and number of solver iterations. This means the accuracy of the simulation is very sensitive to perturbations of the hyperparameters.

Despite this, PBD can be used where physical accuracy is less important than computational speed. For our purposes, it suffices to have PBD procure the correct asymptotic behaviour of the dynamical system. For future work, there are extensions and more recent work on PBD [1, 2, 5] that we can try to adapt to simulate fluid pressures and particle positions.

3. Results

3.1. Modeling hydraulic systems.

A hydraulic system consists of several components, whose formulae are found in [3]. Here we describe the components appearing in Figure 2 and discuss their modeling assumptions with reference to [3]. Here the system schematic consists of a reservoir, fluid, pump, control valves, pipelines and hydraulic cylinder.

A *reservoir* contains a volume of liquid. The hydraulic fluid chosen differs by application but is usually taken to be petroleum or other types of oils. Here, hydraulic fluid is treated as incompressible, but a high-pressure system must account for compressibility.

A *hydraulic pump* is tied to an electric motor to transfer mechanical energy into the system. Given this mechanical part, the pump inlet forms a partial vacuum. This allows atmospheric pressure to push fluid through the inlet and into the pump, which in turn, the pump pushes fluid into the hydraulic system. The pump can be modeled to have constant or variable fluid displacement per cycle. An ideal pump has input torque τ and flow rate Q described by the equations

$$\tau = \frac{D_{\text{pump}}\Delta P}{2\pi}, \qquad Q = D_{\text{pump}} \cdot n.$$

Here D_{pump}, ΔP and n are the pump's volume displacement, pressure differential and shaft speed, respectively.

A *control valve* functions to control pressure, flow rate and direction of a fluid flowing through the hydraulic system. A *directional* control valve controls flow by the area of the orifice for which the fluid moves through. This is modeled by the equation

$$Q_{\text{DV}} = C_D \cdot A \cdot \sqrt{\frac{2\Delta P}{\rho}},$$

where A represents the orifice cross-section area, ΔP is the pressure differential between the input and output of the orifice, ρ is the fluid density, and C_D is a coefficient determined by the shape of the orifice. In simple models, C_D can be treated as constant.

Moving fluid from one component to another is achieved by *pipelines.* Each pipe has a pressure, which is governed by the equation

$$\frac{dP}{dt} = \frac{\beta}{V}(Q_{\text{in}} - Q_{\text{out}}),$$

where P represents the pipe's pressure, β is the fluid's bulk modulus (informally, the resistance to compression), V is the fluid volume of the pipe, and $\Delta Q = Q_{\text{in}} - Q_{\text{out}}$ is the fluid's flow rate differential in the pipe.

Hydraulic cylinders convert the energy stored in the hydraulic system into mechanical energy. As noted before, the cylinder can be used to displace a rigid body. For an ideal cylinder, the 'outgoing' flow corresponds to the rate at which more room is being created in the cylinder chamber for fluid. An increase in pressure pushes the cylinder piston, resulting in more room. This is expressed by

$$Q_{\text{cyl}} = A \cdot \frac{dx}{dt},$$

where A is the area of the piston base pushing the cylinder and x is the displacement of the rigid body. The displacement $x = 0$ corresponds to when the cylinder is fully retracted. Consequently, the cylinder experiences pressure from the flow, which is governed by the equation

$$\frac{dP}{dt} = \frac{\beta}{V_{\text{cyl}}}(Q_{\text{in}} - A \cdot \frac{dx}{dt}).$$

As before, β is the fluid bulk modulus, $V_{\text{cyl}} = A \cdot x$ is the cylinder fluid volume.

3.2. **Model example.** Given the equations of components, we can derive a system of differential equations. Consider the system in Figure 3, which is an ideal cylinder

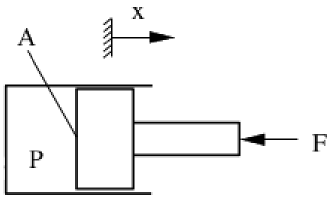


FIGURE 3. A hydraulic system consisting of an ideal cylinder with no flow (image from [3]).

with no incoming flow. This means that $Q = 0$, so the change in pressure of the cylinder is given by

$$\frac{dP}{dt} = -\frac{\beta}{x} \cdot \frac{dx}{dt}.$$

Recall the cylinder piston position is modeled so that $x \geq 0$. Thus if x is decreasing then the pressure is increasing, which matches our intuition. To be able to solve for x and P, we can apply Newton's second law to the piston, so that

$$m\frac{d^2x}{dt^2} = P \cdot A - F.$$

Now we have two equations with two unknowns, x and P, so we can solve this numerically provided we have some initial conditions.

3.3. **Numerical solution with PBD.** Continuing with our Figure 3 example, we took a few different approaches to approximate the system using PBD. One naive approach is to treat the pressures as positions. This does not work well since P and x are described by different physical units. We instead treat pressure and position update steps independently, where in the absence of damping and other modifications in PBD, the algorithm resembles Euler's method. The downside of this is that Euler's method is inaccurate with stiff equations, which worsens from the limit of how small the time step size can be for real-time simulation. Even with our small example, the equations are stiff, since β can be several orders of magnitude larger than other quantities, such as the position, volume or flow.

This is reflected in Figure 4, which plots exact and PBD solutions of the cylinder with no flow. Here $t \in [0, 0.1]$, $\beta = 2.2 \cdot 10^9$, which is the estimated bulk modulus of water, the piston mass is $m = 1$ kilogram, and the base is $A = 0.03^2$ meters squared. The initial conditions are $x(0) = 1$, $x'(0) = 0$ and $P(0) = \beta$. The exact solution in orange is computed using SciPy's `odeint` with very small time step, whereas the PBD solution is computed for a time step of 1200 frames per second (that is, $\Delta t = 0.1/120$). Asymptotically, the pressure appears to be dampening, and as a result the cylinder position and velocity diverge. The asymptotic behaviour is not stable as we predicted.

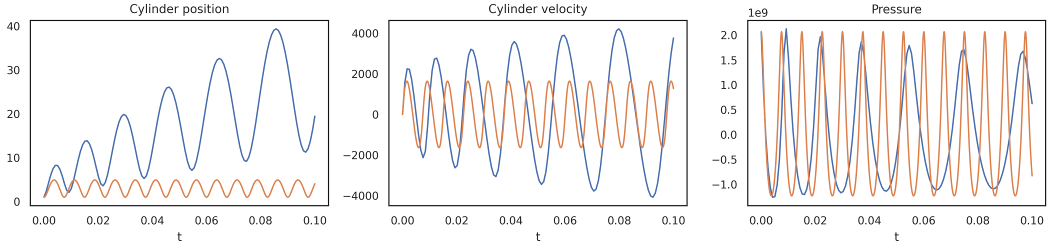


Figure 4. Hydraulic cylinder dynamics from Figure 3 with some set of initial conditions. The orange is the exact solution and blue is a position-based dynamics solution.

4. Conclusion

In this project, we explored the use of position-based dynamics for simulating hydraulic systems. A systemic procedure was demonstrated for obtaining governing equations of simple hydraulic system models. Moreover, PBD was implemented and experimentally tuned to examine the asymptotic behaviour of PBD for hydraulic systems. We hope this work provides a foundation for those actively working on this in the future, that is, developing stable algorithms for real-time simulation of hydraulic systems.

Acknowledgement

We thank the $Math^{Industry}$ organizing committee for providing us a comprehensive, resourceful and interactive workshop for industry career development. For our project, we express gratitude to our industry partner, Serious Labs, and our mentors, Adam Preuss and Slim Ibrahim. They provided us an interesting problem to work on and frequently offered guidance to both enhance our understanding and work through the project.

References

1. Miles Macklin and Matthias Müller, *Position based fluids*, ACM Trans. Graph. **32** (2013), no. 4, 1–12 (en).
2. Miles Macklin, Matthias Müller, and Nuttapong Chentanez, *XPBD: position-based simulation of compliant constrained dynamics*, Proceedings of the 9th International Conference on Motion in Games (Burlingame California), ACM, October 2016, pp. 49–54 (en).
3. Modelon, IDEON Science Park, SE-223 70 LUND, Sweden, *Modeling of hydraulic systems: Tutorial for the hydraulics library*, September 2013.
4. Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff, *Position based dynamics*, Journal of Visual Communication and Image Representation **18** (2007), no. 2, 109–118 (en).
5. Matthias Müller, Miles Macklin, Nuttapong Chentanez, Stefan Jeschke, and Tae-Yong Kim, *Detailed Rigid Body Simulation with Extended Position Based Dynamics*, Computer Graphics Forum **39** (2020), no. 8, 101–112 (en).

Math INDUSTRY

New techniques and technologies in data driven approaches to sustainability *

Symon Islam[†] Sebastián Moraga[‡] Edgar Pacheco[§]
Thomas Pender[¶] Igor Pinheiro[‖]

Abstract

In this paper, we study several aspects of sustainability in agriculture. The need to provide information and projections to farmers and producers is addressed. Then we study several indicators of the environmental impacts on agricultural production.

1 Introduction

Sustainability is the greatest challenge facing the human race, and in the face of global warming coupled with population increase, the strain on vital sectors, like agriculture, is mounting at an accelerated pace. This project is focused on the use of open data to improve understanding, and, ideally, predictability, for the environmental impact from agriculture.

Agriculture is one of the most significant areas of economic activity for countries around the world, and it has a significant environmental impact; it is estimated that agriculture accounts for at least 10% of green house gas emissions in many countries. In nearly every country, there is an impact from agricultural practices through changing land use, water

[*] This research was partially supported by Pacific Institute for the Mathematical Sciences, Mitacs, QuanSight, West Grid and Cybera.

[†] Department of Mathematics, Simon Fraser University, Vancouver, British Columbia, V5A 1S6, Canada. symonyeal@gmail.com.

[‡] Department of Mathematics, Simon Fraser University, Vancouver, British Columbia, V5A 1S6, Canada. smoragas@sfu.ca.

[§] Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, T2N 1N4, Canada. edgar.pachecocastan@ucalgary.ca.

[¶] Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, Alberta, T1K 3M4, Canada. thomas.pender@uleth.ca.

[‖] Department of Mathematics, University of British Columbia, Vancouver, British Columbia, V6T 1Z2, Canada. igor.ivpp@math.ubc.ca.

consumption, use of fertilizer, GHGs and other emissions. From a local perspective agriculture creates different dynamics for indigenous plant and animal life in addition to creating different micro-climates. On a more broad scale, agriculture can put pressure on entire river systems and lead to changing weather patterns as atmospheric humidity and solar radiation emissions change.

The ideal outcome of this project is the design of a model and a system of action for environmental impacts that leverages open data and provides a complement to the current TheoryMesh system which is capturing operational impacts from farm activities. Invariably, the project and its outcome will be data driven.

There are many data sources available to investigate environmental changes and impacts due to agriculture. Furthermore, there is a proliferation of data sets that convey information about practices, agricultural production, and green house gas emissions. Combining data across data sources and interpreting the data in new ways could provide better insights on sustainability. Using machine learning techniques to creating models to describe these impacts could improve planning and shift practices to reflect longer term environmental impact. For example, technologies like Blockchain may provide a foundation to create an immutable data ledger for environmental impact while also leveraging smart contracts to take action on data when conditions are met.

For a broader view, the reader is encouraged to read The United Nations Sustainable Development Goals that describe a multi-faceted view of sustainability covering environment, economic and societal factors [6].

Finally, we note that tables and figures are reserved for an Appendix at the end of this document.

2 Models and results

In this section we discuss briefly the models used on the data collected in this research. The models we implemented are simple linear regresion models of the type

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{i,j}, \tag{2.1}$$

for the observations $i \in \{1, 2, \ldots, N\}$, where N are total number of observations, and where there are k dependent variables .

Furthermore, in this study we use the dependence between two specific quantities using a correlation coefficient matrix, with the Pearson's correlation coefficients. This method is widely used to see linear correlation between variables that also uses a least squares fitting to the data obtained.

3 Problems

The issues alluded to above are broad, and they are difficult to quantify and to analyze; indeed, agricultural sustainability is a function of a large number of variables. In an effort to make the problem more tractable, we focused on two smaller problems.

The first problem arose in the following way. There is a lot of work being done on the broader, macro scale. However, it is interesting that not much work has been done on the local, micro level. For instance, a common complaint among contemporary farmers is that information cannot be compiled and disseminated in the same ways anymore. It used to be the case that a producer might say "Plant this crop on this day because that is what has always worked." This is no longer tenable with the changing climate and increased weather volatility.

All this is to say that the productivity of agricultural fields, what producers depend on, is being negatively affected, and new approaches to collecting, interpreting, and disseminating relavent information need to be produced.

As a second problem, we see that environment problems have been a real matter in the last century. Factors that could be correlated, for example greenhouse gases emission, humidity, or heat, can explain why certain aspects of area of production can affect others. The main issue here is try to identify such factors in Canada and the impact they may have on the environment and the surroundings if one or more of them change.

3.1 Productivity of Canadian Farms

There are many areas in the world that are experiencing the catastrophic effects of climate change. In the Middle East, many of the farmers who have had to rely on local tributaries to sustain their crops are having to consider different ways of life as water levels of these rivers and lakes decrease and, ultimately, disappear. As these climate events unfold, the UN [5] warns of a proliferation of conflict over the increase in water scarcity.

For more northern regions, like Canada, the situation may develop differently. As the climate warms, areas to the north will continue to become more viable for a greater variety of agricultural practices and commodities. It becomes important, then, to understand how climate variables can effect changes in crop yields.

For instance, consider the number of frost free days in the province of Alberta[1]. In each climate model, where we note that rcp26, rcp45, and rcp85 are, respectively, the low, mid, and high emisions models, and as indicated by Figure 1, the number of consecutive frost free days will invariably increase. With the increased length of the growing season, there will be more opportunities to grow a greater variety of crops in regions that have been hitherto difficult to cultivate.

There are other climate variables to consider as well. Observe the cumulative monthly mean temperature in Alberta (sum of monthly mean temperatures) shown in Figure 2 in addition to the consecutive frost free days [1].

To evince that the productivity of Canadian farms is, in fact, increasing, and to provide specific crop data per locale for producers, we can consider data like that shown in Figure 3, which shows the change in barley yields in Alberta [4]. In this figure, we can clearly see an upward trend in the relative yields of barley in Alberta.

Understanding the increase in production, we can then ask which climate variables can be correlated to farm productivity. These climate projections per locale can be used by producers in order to project at what points they may profitably consider changes in the location of their fields and the type of crops they choose to grow in these locations. Continuing to consider barley yields, we can compare historical yield data to the climate models in order to give a simple projection of future yields.

Assuming the yield to be a function of time and a number climate variables, say, the number of days with max temperature greater than 32 °C, the number of frost free days, the cumulative precipitation, and the cumulative monthly mean temperature, we endeavour to use the usual multilinear regression techniques and tools to exterpolate into the future (here we have used the popular sklearn module). We note the projections here are only illustrative and hint at over-all trends.

Now, as is clear from Figure 4, in all but the worst case emissions scenario, barley yields in Alberta are projected to continue to grow. All this confirms the suspicion that climes like that found in Canada will become more and more important in crop production as the climiate continues to change.

3.2 Indicators of Environmental Impact

In this work, we first stablished the main problems and overall challenges that food process and agriculture may face in the future. The main results showed in Section 2, asserting the existence of a linear correlation between variables from food production. With this in mind we have a better understanding of the indictors for sustainability. Where the preliminary results show why in certain cases, while the poultry production increases per capita some other factors decrease, like beef or pork production. The results of this research are shown in figures 5, 6, 7, 8, 9 and 10.

The previous research indicates there are many other factors to consider, and this is the starting of an opportunity for more investigation about the challenges of food production.

4 Conclusions

Next steps to consider are things such as the following: To look at the sustainability of the economic activities in Canada, and to develop a system to tackle problems as land use, water consumption and use of fertilizers.

The preliminary results obtained in the work are promising, as well the need for further improvements to open problems such as certification for the food suppliers and traceability of the products. These results could be reported in a future work.

References

[1] Climate Data Canada Climate Data for a Resilient Canada
DOI: https://climatedata.ca/

[2] D. McInnes, Agri-food sustainability targets. A selected overview. DMci Strategies, October 2003.

[3] K. Parris, Sustainable management of water resources in agriculture. OECD Publishing. France, 2010.

[4] Statistics Canada Table 32-10-0359-01 Estimated areas, yield, production, average farm price and total farm value of principal field crops, in metric and imperial units
DOI: https://doi.org/10.25318/3210035901-eng

[5] United Nations ClimateChange
DOI: https://www.un.org/climatechange?gclid=CjwKCAjwvuGJBhB1EiwACU1AidbTCITKc5CBRuumT

[6] United Nations THE 17 GOALS — Sustainable Development
DOI: https://sdgs.un.org/goals

Appendix

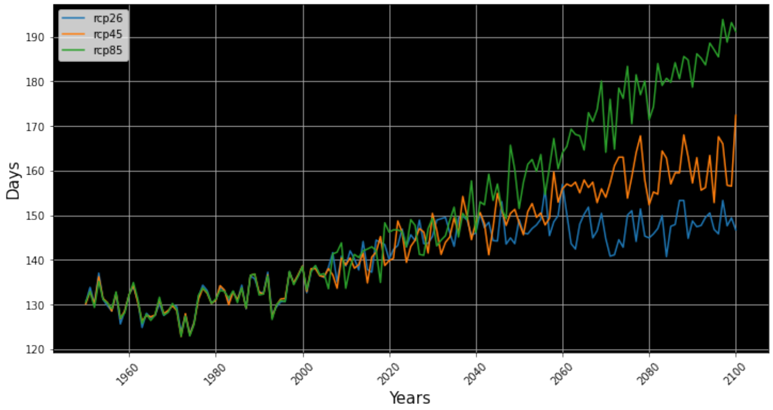


Figure 1: Consecutive Frost Free Days in Alberta

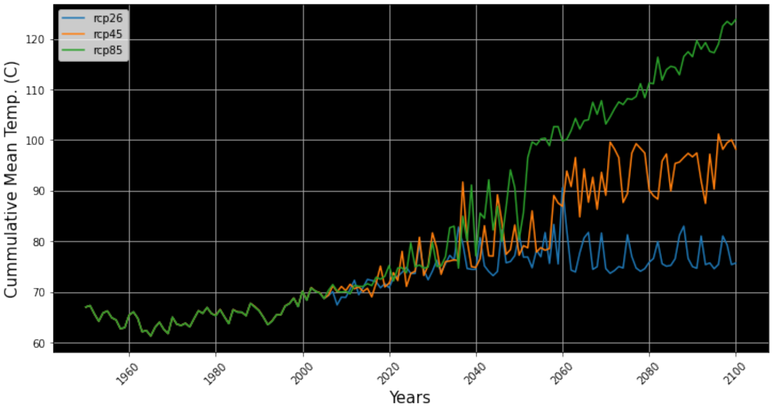


Figure 2: cumulative Mean Temperatures in Alberta

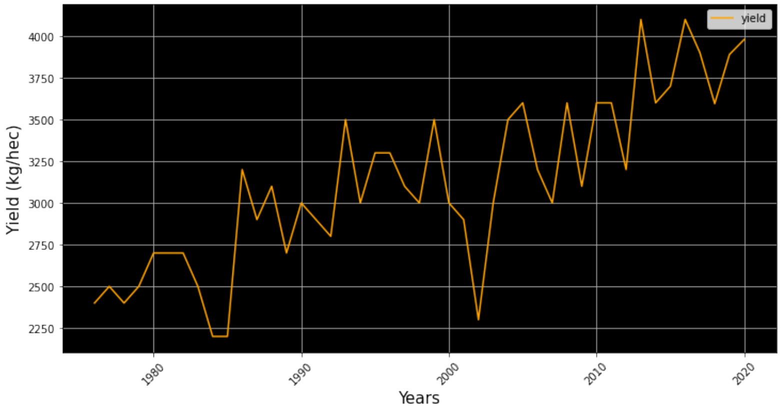


Figure 3: Barley Yields in Alberta

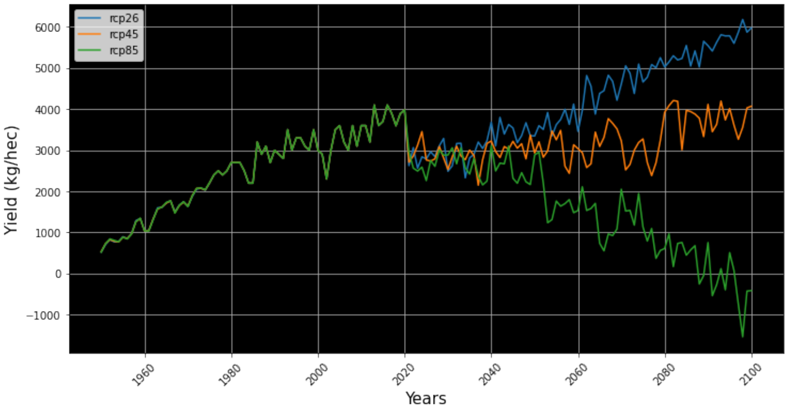


Figure 4: Projected Barley Yields in Alberta

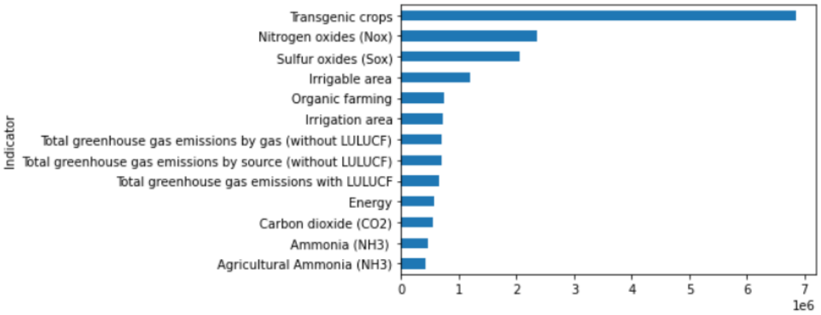


Figure 5: Comparison of the different types of indicators values in Canada, from 1984 to 2017. Shows only the indicators over the mean value, where transgenic crops was the most used indicator.

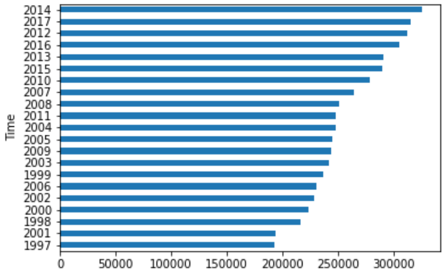
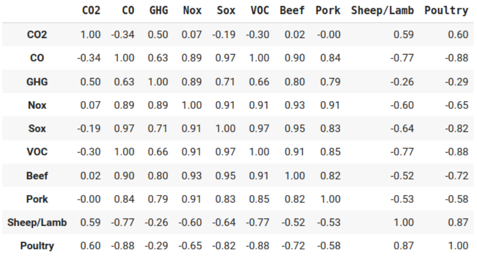


Figure 6: Shows the amount of indicators per capita from 1984 to 2017, starting around 2000000 to near 350000 units per capita. The year where more indicators were used was 2014 and 2017. Besides, the results shown are only the ones above the average within this time frame.



	CO2	CO	GHG	Nox	Sox	VOC	Beef	Pork	Sheep/Lamb	Poultry
CO2	1.00	-0.34	0.50	0.07	-0.19	-0.30	0.02	-0.00	0.59	0.60
CO	-0.34	1.00	0.63	0.89	0.97	1.00	0.90	0.84	-0.77	-0.88
GHG	0.50	0.63	1.00	0.89	0.71	0.66	0.80	0.79	-0.26	-0.29
Nox	0.07	0.89	0.89	1.00	0.91	0.91	0.93	0.91	-0.60	-0.65
Sox	-0.19	0.97	0.71	0.91	1.00	0.97	0.95	0.83	-0.64	-0.82
VOC	-0.30	1.00	0.66	0.91	0.97	1.00	0.91	0.85	-0.77	-0.88
Beef	0.02	0.90	0.80	0.93	0.95	0.91	1.00	0.82	-0.52	-0.72
Pork	-0.00	0.84	0.79	0.91	0.83	0.85	0.82	1.00	-0.53	-0.58
Sheep/Lamb	0.59	-0.77	-0.26	-0.60	-0.64	-0.77	-0.52	-0.53	1.00	0.87
Poultry	0.60	-0.88	-0.29	-0.65	-0.82	-0.88	-0.72	-0.58	0.87	1.00

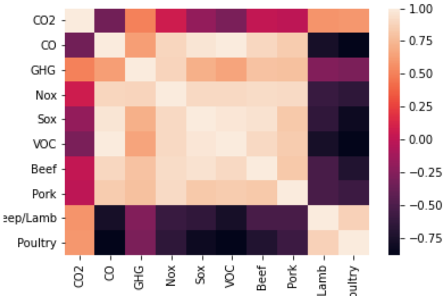


Figure 7: We compare different factors from environment impact with a correlation matrix. (**Left**) Shows the pairwise correlation coefficients between various variables of the Canadian food production, such as GHG and farm supplies. (**Right**) Shows a visualization of the corresponding correlation matrix on the left.

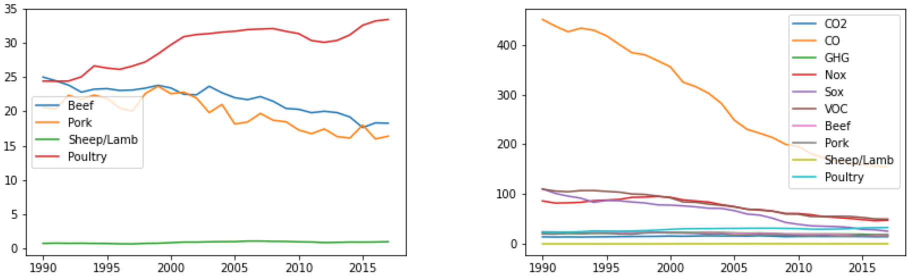


Figure 8: Comparison in time of food supplies production per-capita from 1990 to 2017. (**Left**) Shows the evolution of different products from farms, whereas we observe an increasing number of poultry but a decreasing number of other supplies as beef and pork. (**Right**) Shows the evolution of the farm products along GHG produced in the same time frame.

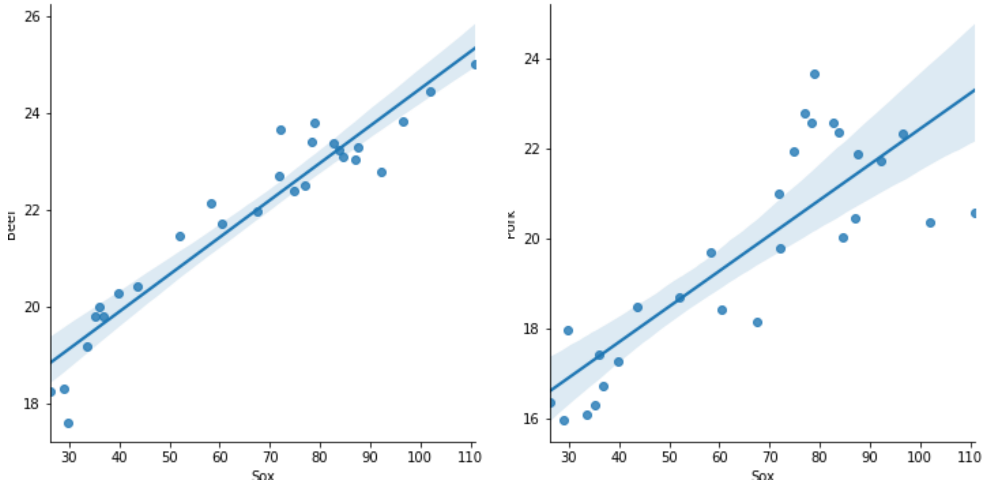


Figure 9: Shows a best-fit line for two variables and the respective confidence intervals. (**Left**) Comparison for a linear correlation between beef production per-capita and Sox gas emited. (**Right**) Comparison for a linear correlation between pork production per-capita and Sox gas emited.

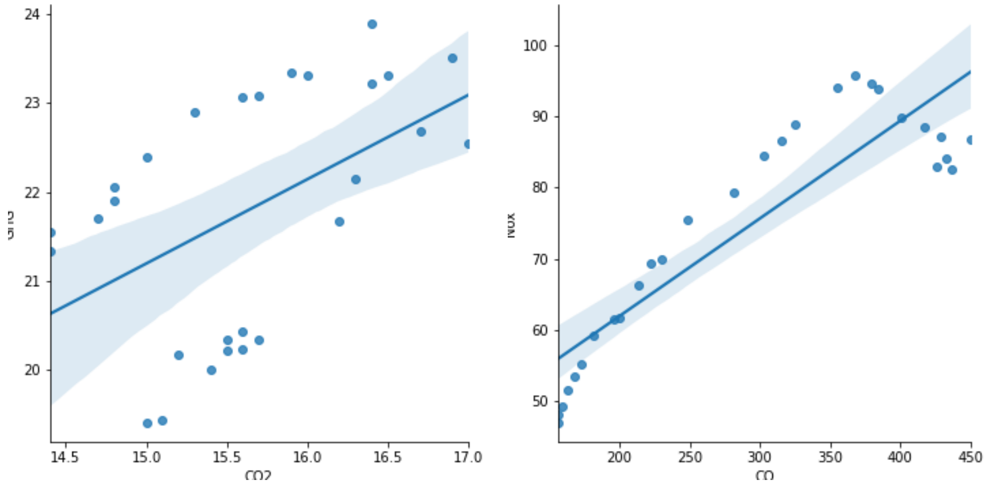


Figure 10: Shows a best-fit line for GHG and one specific gas with the respective confidence intervals. (**Left**) Comparison for a linear correlation between GHG production per-capita and CO2 emited. (**Right**) Comparison for a linear correlation between GHG production per-capita and CO emited.