SIE 512 Final Project
Spatial Analysis of Albertan Wildfire Data

Chris Ewanik

Abstract[1]

This paper outlines an analysis of the spatial distribution, trends and prediction of Albertan Wildfire data from 1930 onwards. Fire counts suggest that fires are increasing and changing their seasonal pattern, consistent with research from other studies. The data exhibits spatial clustering according to quadrat tests and nearest neighbour probabilities. Artificial neural networks correctly classified the size of a fire (large or small) 93% of the time. The dataset is limited by its lack of national standards and missing data. Further tests should be done that continue this work incorporating other more meaningful data sources that would provide a model that could be used closer to real-time.

1. Introduction

The government of Canada provides a few remarkable facts about Canadian Wildfires. In Canada, an average of 2.5 million hectares of forest burn annually. While only 3% of all wildland fires get to a size larger than 200 hectares, those mega-fires are responsible for burning 97% of the total area. Finally, the cost of fighting fires ranged from $800 million to $1.5 billion a year in the last decade (Government Of Canada, 2022). Additionally, these wildfires emit a considerable amount of carbon. The addition of carbon from fires and the subtraction of plant life brings more carbon into the atmosphere.

Wildfires are highly destructive to property. The wildfires of Slave Lake and Fort McMurray, Alberta, are two expensive and disruptive fires. They each took away a tremendous amount of homes with substantial insurance costs and costs on quality of life. Finally, wildfires are highly dangerous to Canada's outstanding wildlife. The fire and habitat loss harm many precious wildlife species, particularly in National parks.

With all of the above reasons in mind understanding the distribution and trend of wildfires is important to many stakeholders. Additionally, prediction models could benefit society with increased home security, reduced taxes and reduced emissions. This model uses spatial and temporal features to classify, trend and predict wildfires in the province of Alberta through density functions, spatial statistics and artificial neural networks.

[1]

2. Related Work

New research has recently been published on wildfire data. One work of particular interest is "Spatial and temporal pattern of wildfires in California from 2000 to 2019". This paper looked at wildfires in California from 2000-2019 and compared them to fires between 1920-1999 (Li & Banerjee, 2021). The researchers specifically looked at the probability density distribution of wildfires, temporal distribution trend, spatial distribution characteristics, and the correlation and importance of explanatory natural and social variables with wildfires. The study found that the frequency of fires has increased, specifically small, human-caused, and natural fires. However, the rate of large fires has remained constant. The study continued to find that although the rate of large fires remained constant, large fires now burn more areas than in the previous century. Finally, more regions were experiencing fires compared to the past. The current paper draws inspiration from this work by testing the distribution and complete spatial randomness by decade. This paper will also make distinctions between large and small fires.

The paper "Predictive Modelling of Wildfires: A new dataset and machine learning approach" was another relevant piece of literature published. This paper used support vector machines and artificial neural networks on satellite images to predict the occurrence of wildfires (Sayada, Mousannifb, & Moatassimea, 2019). The researchers combined crop data, meteorological conditions, and fire indicators in a "Big Data" format on Databricks to create a highly predictive accuracy of 98.32%. Additionally, this paper used data from the Canadian Wildland Fire Information System, which is directly related to the data published in this paper. This paper attempts to use artificial neural networks to classify the size of a fire (large or small).

3. Dataset and Features

The data used in this paper is from the Canadian National Fire Database (Natural Resources Canada, 2022). Specifically, data is ignition data for Canadian Wildfires from 1930 on. Data early is sporadic and inconsistent. The dataset, in total, holds 423381 observations of 27 variables. The dataset includes ignition longitude and latitude and polygon data for fire perimeters. There are also more tabular data types, such as fire cause and eco-district. Figure 1 shows a plot of all Canadian ignitions stored in the data set.

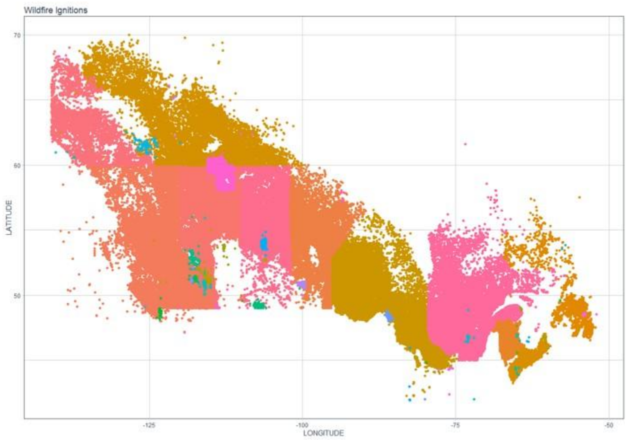

Figure 1

While large, the dataset can hold potential errors through location approximation, unmapped fires, and different mapping techniques. A challenge of this dataset is the lack of true national standards. For example, this study is concerned with the Fire Type variable. Since this encoding is not standardized or documented, interpretation for non-domain experts is difficult. Even if models reveal these variables are significant, domain experts need to be consulted to understand the meaning.

Although there is data for all provinces and territories, this study focuses on Alberta. This was chosen because Canada's vast size and diversity were too large of an area for the scope of this work. Using Alberta would yield more meaningful results.

For the spatial analysis section, the decades of 1970, 1990 and 2010 are all used to compare over time. These datasets were chosen as reporting before the 1970s suffered from missing data. Due to the ambiguity and a need for a national data standard, there are potentially (likely) some data issues present for the 1970s (and all decades). Still, these three would yield a comparison at least indicative of real representations, particularly for ignitions believed to be relatively accurate.

A variety of data preprocessing steps were used to prepare for deep learning. First, the package \`elevatr\` was used to extract each ignition's elevation (in meters). \`Elevatr\` uses an API to connect to Amazon Web Service Terrain tiles to pull the height. Next, the package \`Recipes\` is used to transform the dataset into a format better for machine learning. Discretized bins were constructed for latitude and longitude, nominal variables were dummy encoded, and interaction effects were mutated for combinations between \`fire type\`, \`cause\` and \`ecozone\`. The training dataset contains 33369 observations with 246 columns, with the testing set holding 11123 observations also with 246 columns.

4. Methods

This study uses a variety of approaches to model and assess the trend of Canadian Wildfires. Firstly, a visual exploratory data analysis is done to determine the effect of different independent variables like ecology zone, year and month. Next, a point pattern analysis algorithm assesses the influence and potential change in yearly trends and seasonality on large and small fires. Finally, an artificial neural network is constructed to classify large and small fires. Large fires are any fire equal to or greater than 200 hectares.

Various point pattern analysis techniques are used to assess spatial and spatiotemporal interaction for the 1970s, 1990s and the 2010s. A simple density function is used to plot the intensities of fire ignitions. Next, a Chi-Squared test of Complete Spatial Randomness using quadrat counts is conducted on a 5x5 grid to test the null hypothesis of complete spatial randomness. Three separate point process models are fit to the data: Homogeneous Poisson, First Order Inhomogeneous Poisson and Second Order Inhomogeneous Poisson. These models are compared with ANOVA and AIC to assess the best fit. While the quadrat test acts sufficiently to test the null hypothesis of complete spatial randomness, the accuracy of the test is dependent on the size of the grid chosen. Therefore the nearest neighbour distance probability function is fitted to provide another test of the null hypothesis of complete spatial randomness. Finally, a Mantel test is implemented to test space-time interaction using space and time decisions. The Mantel test tests the null hypothesis that spatial distances between cases are independent of the time distance between those cases (Long, 2022).

H2O is an open-source, distributed, in-memory machine-learning platform that utilizes gradient-boosted models, generalized linear models, deep learning and more (H2O.ai, 2022). This platform allows for tremendously quick training and interpretable insights. In this study, the automated machine learning function was restricted to training only deep learning models. This caused h2o to train models with different epochs, dropout rates, regularization values, learning rates and more, similar to a grid search.

5. Experiments/Results/Discussion

Figures 2 and 3 show the count of fires by year and month. While the early years (before 1970) suffer from a lack of data, it is still clear that the count of large and total fires is rising. Large fires, in particular, have been increasing for the last two decades.

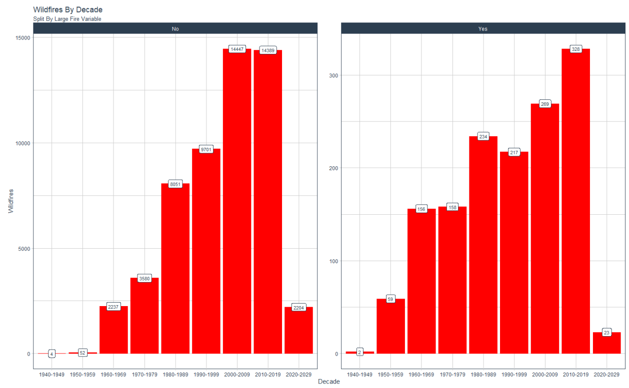

Figure 2

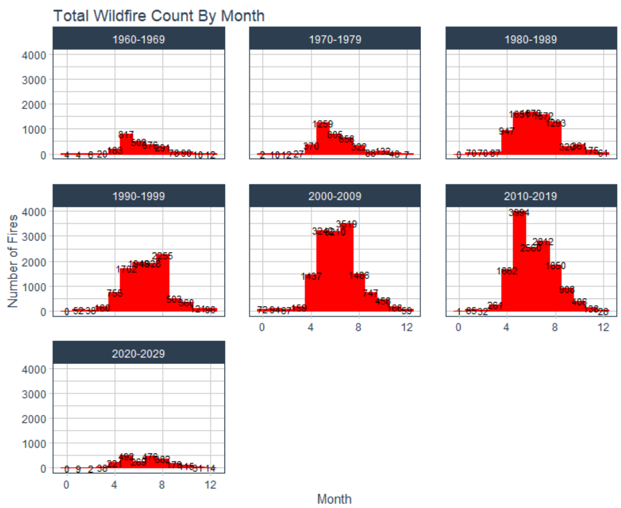

Figure 3

Figure 3 shows the monthly count of wildfires split by the decade. While these plots show the aforementioned increase in count, it also appears that much of this rise is due to fires starting earlier in the year. The subplot of Figure 3 for Wildfires in 2010-2019 shows a significant increase in May. These results would be consistent with other research suggesting that fire seasons are starting earlier than before (Li & Banerjee, 2021).

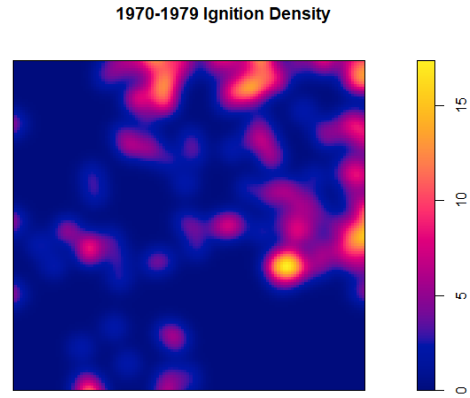

Figure 4

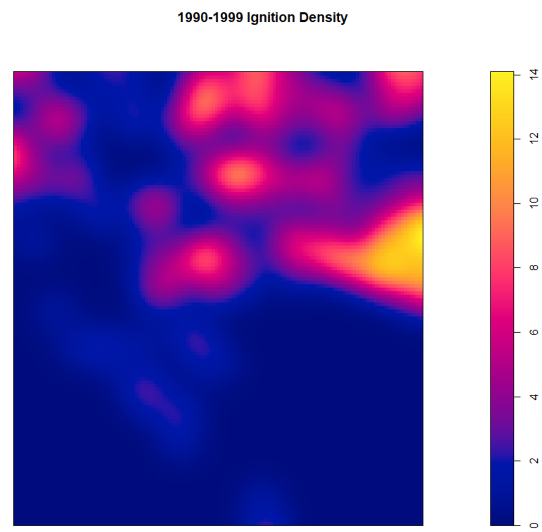

Figure 5

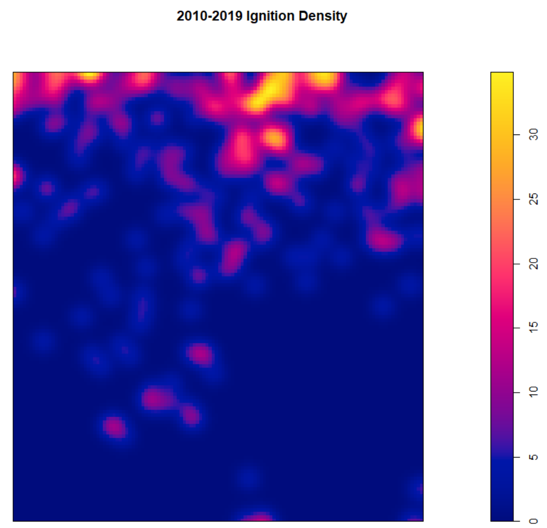

Figure 6

Figures 4, 5, and 6 all show the plotted densities. Areas of high intensity are frequent in the North and East parts of the province, with the south having fewer fires. One can infer that this reason is mainly ecological and logistical. Southern Alberta is essentially prairie and agricultural land. Therefore, having no trees makes it impossible to have a forest fire. Additionally, northern Alberta is more remote and difficult to access. Consequently, it is much more challenging to detect fires and for crews to fight them.

Decade	P-Value	Null (CSR)
1970-1979	2.1e-12	Reject
1990-1999	2.2e-16	Reject
2010-2019	2.2e-16	Reject

Table 1

Table 1 shows the quadrat test results. For this test, all three decades reject the null of complete spatial randomness. This was to be expected from the examination of the spatial densities.

Decade	Lowest AIC	Model
1970-1979	55.2	First Order Inhomogeneous Poisson
1990-1999	-36.0	Second Order Inhomogeneous Poisson
2010-2019	-437.9	Second Order Inhomogeneous Poisson

Table 2

Table 2 shows the results of spatial model comparisons. Interestingly, The 1970s were best modelled by a first order inhomogeneous Poisson process, and the other two were best modelled by a second order inhomogeneous Poisson. Further tests should be done on other decades to test whether 1970 was an outlier supporting a different model or if a fundamental change to spatial characteristics changed, causing a difference in the model.

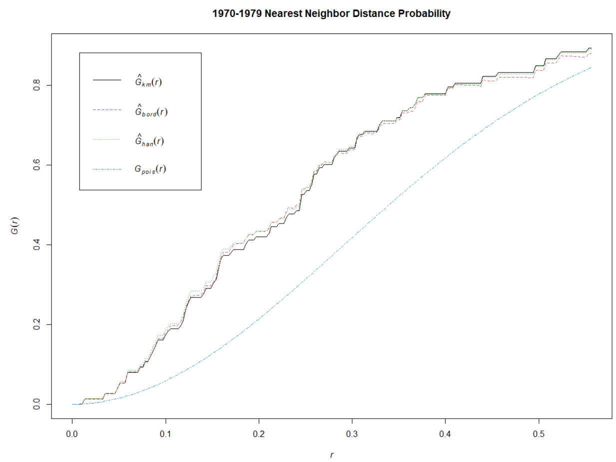

Figure 7

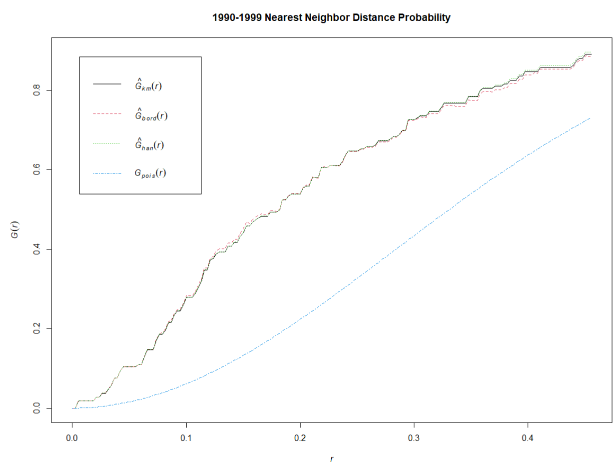

Figure 8

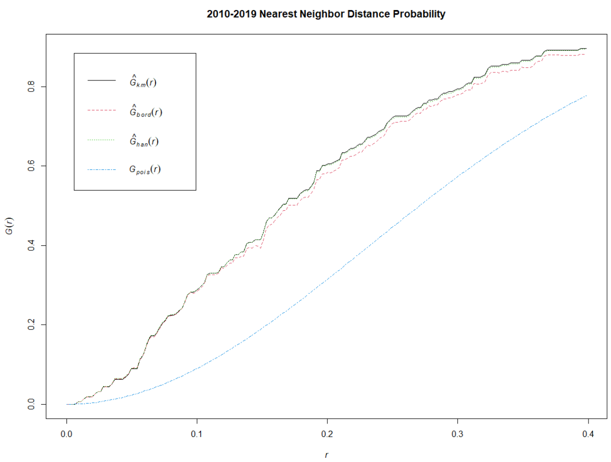

Figure 9

Next, figures 7,8 and 9 all show the G functions testing spatial clustering. All three decades show G functions above the line indicating spatial clustering (consistent with the results of the quadrat test).

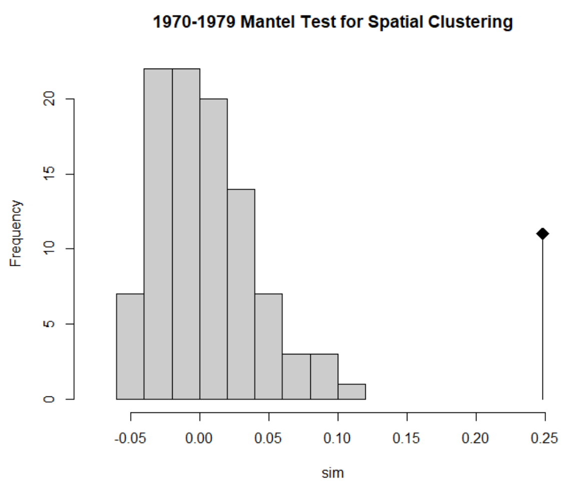

Figure 10

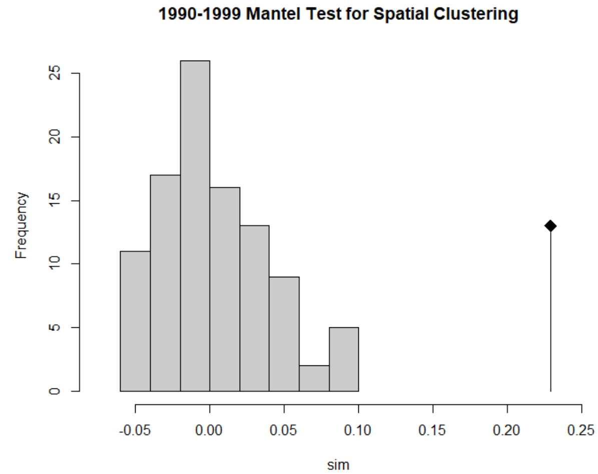

Figure 11

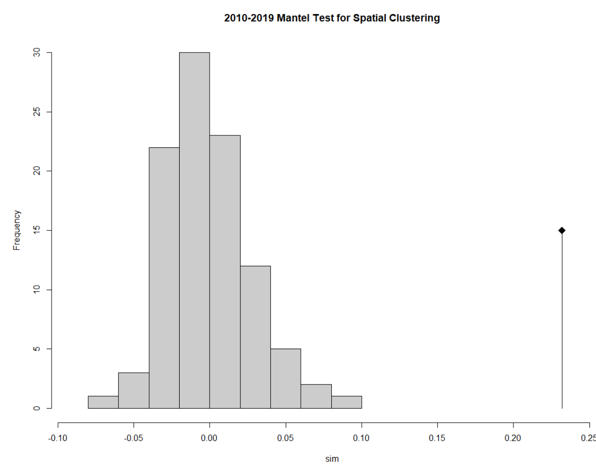

Figure 12

Decade	P-Value	Null (No Space Time Interaction)
1970-1979	0.01	Reject
1990-1999	0.01	Reject
2010-2019	0.01	Reject

Table 3

Figures 10, 11 and 12, along with table 3, show the results of Mandel's Test. All three plots and the table show that each null hypothesis of no space-time clustering is rejected.

The neural network in this study had many of its hyperparameters tuned by h2o. Some parameters were chosen, such as a stopping metric of AUCPR (Area Under the Precision-Recall Curve) and class balancing to address the lack of large fires in the dataset. Five fold cross validation was also implemented to fight overfitting, which was essential with many independent variables. The automated machine learning algorithm determined that an input dropout ratio of 0.2, a hidden dropout ratio of 0.4 and a rectifier with a dropout activation function were optimal (likely due to the high number of parameters).

Large Fire	Classification	Rate
0	0.938	10248/10925
1	0.45	89/198
Totals	0.93	10337/11123

Table 4

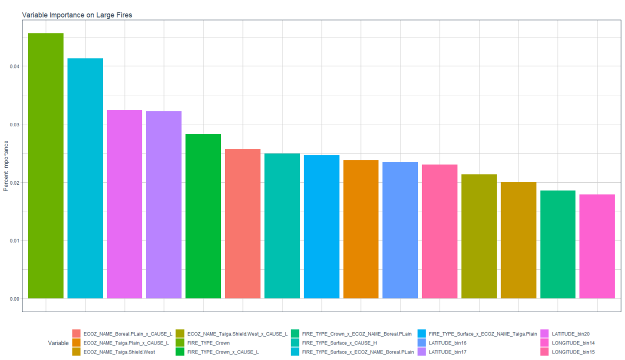

Figure 13

Finally, table 4 shows the classification results of the neural network. Overall the neural network achieves a high level of accuracy, predicting the binary fire size correctly 93% of the time. Additionally, figure 13 shows the variable importance to the neural network. The model prioritizes the fire type, the interaction of lightning, certain ecological zones and certain latitude and longitude bins.

6. **Conclusion and Future Work**

Overall, the work of this study yielded a variety of exciting results. Fires were shown to have high densities in northern and eastern Alberta, with relatively few fires in the southeastern parts of the province. Fires have increased over time, and there is reason to believe that seasonal patterns have changed. The results of the quadrat test and the nearest neighbour probabilities indicate confidently reject the null hypothesis of complete spatial randomness. Different decades were better fit by other spatial models, as the 1990s and 2010s were best represented by a second order inhomogeneous Poisson model, while the 1970s best fit a first-order inhomogeneous Poisson. Finally, an artificial neural network built with h2o was shown to correctly classify 93 % of fires (large or small).

While this study has many significant findings, there are limitations. The differences in yearly and seasonal trends appear obvious but should be tested for statistical significance. T-tests would be an obvious next step in determining whether or not the discrepancy was significant. The fitted models showed a difference in the decades. Further tests should be done to see what models other decades support, and a deeper dive into understanding why could help identify pattern changes. While the Neural Network had a high classification accuracy level, it only correctly predicted 45% of large fires. These are the fires that one would want a model to predict accurately, as they cause a large amount of damage.

Additionally, while this dataset lends itself to classification, these variables are collected after the fire. This means this model is not possible in a real-time version, which is likely the best version. Additionally, this analysis did not include raster data like meteorological data that has aided the predictive accuracy of neural networks (Sayada, Mousannifb, & Moatassimea, 2019). Incorporating more data sources into this model could provide a more confident, scaleable and reliable model. The lack of a true national standard makes the interpretation of the model complex. A more widely accepted and clear criteria would help outside analysts contribute to these projects. The addition of time series modelling could also support this project. Mantel's test showed that there is statistical evidence for space-time interaction. Since wildfires need fuel to burn, it seems highly unlikely that fires (particularly large fires) would repeat themselves in the same locations. A time series-based model that could calculate risk based on the last fire could help this report. Finally, other forms of point pattern analysis could be helpful. A bivariate point pattern analysis would be especially useful if it were formatted so that one could test where fires did and did not happen. This would allow for tests that predict whether the fire will be large and whether it will happen at all.

References

Government Of Canada. (2022, 04 14). *Forest Fires*. Retrieved from Government of Canada Natural Resources: https://www.nrcan.gc.ca/our-natural-resources/forests/wildland-fires-insects-disturbances/forest-fires/13143

H2O.ai. (2022, 12 13). *H2O Overview*. Retrieved from H2O.ai: https://h2o.ai/platform/ai-cloud/make/h2o/

Li, S., & Banerjee, T. (2021). Spatial and temporal pattern of wildfires in California from 2000 to 2019. *Scientific Reports*.

Long, A. (2022, 12 13). *Mantel's Test*. Retrieved from NKU Edu: https://www.nku.edu/~longa/geomed/stats/Mantel/Mantel.html

Natural Resources Canada. (2022, 12 13). *CWFIS Datamart*. Retrieved from Natural Resources Canada: https://cwfis.cfs.nrcan.gc.ca/datamart/metadata/nfdbpnt

Sayada, Y. O., Mousannifb, H., & Moatassimea, H. A. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal*, 130-146.