# ESTIMATING DWELLING DISTRIBUTION IN RURAL ALBERTA

# “THE LIGHTS ARE ON BUT IS ANYONE HOME?”: ESTIMATING DWELLING DISTRIBUTION IN RURAL ALBERTA

By SAMI KURANI, H.B.Sc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2020) Hamilton, Ontario (Geography)

TITLE: “The lights are on, but is anyone home?”: Estimating dwelling distribution in rural Alberta

AUTHOR: Sami Kurani, H.B.Sc. (McMaster University)

SUPERVISOR: N. Yiannakoulias

NUMBER OF PAGES: x, 85

# ABSTRACT

With Canada's increasing population, natural disasters such as flooding events will have an increasing impact on human populations. The severity of these events requires that decision makers have a clear understanding of the flood risks that communities face in order to plan for and mitigate flood risks. One key component to understanding flood risk is flood exposure, an element of which is the presence of structures (e.g., residences, businesses, and other buildings) in an area that could be damaged by flooding. Presently, several resources exist at both the national and global level that can be used to estimate the spatial distribution of structures. These resources are typically generated at global scales and do not account for regional or local data or processes that could enhance the accuracy and precision of exposure estimation in sparsely populated areas. The present study investigates the feasibility of creating a region-specific dwelling distribution model that helps improve estimation of residential structures in rural areas. Herein, we describe a rural dwelling distribution model for the province of Alberta that can be used to assist in the estimation of structural exposure to flood risk. The model is based on a random forest classification algorithm and several publicly available datasets associated with dwelling and population density. The model was validated using visually referenced data collected from earth imagery. The resulting dwelling layer was then evaluated in its ability to spatially disaggregate census dwelling counts, as well as predict dwelling exposure in several scenarios. This method appears to be a useful alternative to globally scaled models, or using the census alone, particularly for rural areas of Canada.

# ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Dr. Niko Yiannakoulias. Throughout both my undergraduate and graduate degrees, he has been an incredible mentor, and without his support, patience, and guidance, this thesis would not be possible. Over the past three and a half years, working with the Health Geomatics lab at McMaster University has been such an incredible experience, both challenging and shaping me, and continually pushing me out of my comfort zone.

Next, I would like to thank my friends, former lab mates, Shelby Sturrock, Julien Gordon, and Connor Darlington, as well as my current lab mate, Cathy Slavik, for dealing with me these past two and a half years while I would bore them with population maps, complain about datasets not playing with each other, and other general research mishaps. Whether it be taking a night off from research or being there to hear me vent on just about anything, I appreciate and thank all of you.

Lastly, I need to thank my family for their love and support during these (unprecedented) times. They were especially supportive and loving during these last few months of writing, which along with the events taking place in the world, was sorely needed, either in the form of a home-cooked care package, or a well-timed hug.

## TABLE OF CONTENTS

### CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND ........ 1
  1.1.1 Introduction ........ 1
  1.1.2 Measuring Dwelling Exposure ........ 2
1.2 RESEARCH OBJECTIVES ........ 3
1.3 CHAPTER OUTLINE ........ 3
  1.3.1 Summary ........ 3
  1.3.2 Chapter 2 ........ 4
  1.3.3 Chapter 3 ........ 4
  1.3.4 Chapter 4 ........ 4
  REFERENCES ........ 5

### CHAPTER TWO: CREATING A RURAL DWELLING DISTRIBUTION MODEL FOR THE PROVINCE OF ALBERTA

2.1 INTRODUCTION ........ 6
  2.1.1 Background ........ 6
  2.1.2 Measuring Dwelling Exposure ........ 7
  2.1.3 Machine Learning and Population Mapping ........ 10
  2.1.4 The Current Challenge ........ 12
  2.1.5 Objective ........ 13
2.2 METHODS ........ 14
  2.2.1 Study Area ........ 14
  2.2.2 Grid layer: ........ 15
  2.2.3 Data ........ 16
    2.2.3.1 Raster Data ........ 17
    2.2.3.2 Vector Data ........ 18
  2.2.4 Data Collection via Photointerpretation: ........ 23
  2.2.5 Modelling: ........ 25
  2.2.6 Validation and Study Area Application: ........ 29
2.3 RESULTS ........ 30

2.3.1 Photointerpretation of Quarter Sections: 30
2.3.2 Descriptive Statistics of Ancillary Spatial Data: 30
2.3.3 Model Output: 31
2.3.4 Scaling the Models to the Study Area 38
2.4 DISCUSSION 43
2.4.1 Summary of Findings 43
2.4.2 Limitations 44
2.4.2.1 Nighttime Light Data 45
2.4.2.2 Distinguishing between industrial and residential sites 46
2.4.2.3 Data Collection & Random Forest Modelling 47
2.5 CONCLUSION 47
REFERENCES 49

CHAPTER THREE: EVALUATING REGIONAL DWELLING MODEL PERFORMANCE IN COMPARISON TO GLOBAL HUMAN SETTLEMENT LAYERS

3.1 INTRODUCTION 54
3.1.1 Background 54
3.1.2 Evaluating Human Settlement Layers 55
3.1.3 The Rural Dwelling Distribution Model 57
3.1.5 Objective 57
3.2 METHODS 58
3.2.1 Study Area 58
3.2.2 Data 60
3.2.3 Human Settlement Layers 61
3.2.4 Evaluating the Human Settlement Layers 64
3.3 RESULTS 67
3.3.1 Evaluation 1 – Dwelling Count Allocation Performance 67
3.3.2 Evaluation 2 – Scenario Sites 74
3.4 DISCUSSION 75
3.4.1 Dwelling Count Allocation 76
3.4.2 Scenario Site Evaluation 77
3.4.3 Limitations 78

3.5 CONCLUSION ........ 79
REFERENCES ........ 80
Appendix A ........ 82

CHAPTER FOUR: CONCLUSION

4.1 INTRODUCTION ........ 83
4.1.1 Summary of Major Findings ........ 83
4.1.1 Chapter Two ........ 83
4.1.2 Chapter Three ........ 84
4.2 CONCLUSIONS ........ 84

# LIST OF TABLES AND FIGURES

## TABLES

Table 1: Explanatory Variables ........ 20
Table 2: Recorded Dwelling Frequencies........ 30
Table 3: Descriptive Statistics of Key Predictors ........ 31
Table 4: Regression Model Results ........ 35
Table 5: Accuracy Table from Binary Approaches ........ 37
Table 6: Specificity & Sensitivity Table from Binary Approaches........ 37
Table 7: Ranked Nighttime Light Index Categories ........ 63
Table 8: Dwelling Count Allocation Results (Total % of each Layer)........ 72
Table 9: Human Settlement Layer Mean Absolute Error ........ 73
Table 10: Dwelling Exposure by HSL (% exposed dwellings with respect to scenario site)....... 75

## FIGURES

Figure 1: Administrative boundaries compared to a gridded surface. ........ 8
Figure 2: Study area extent ........ 15
Figure 3: The Dominion Land Survey ........ 16
Figure 4: From left to right - The entire province, the study area (No overlap with Towns / Urban Centers / Cities and waterbodies), and the randomly selected cells. ........ 23
Figure 5: An example of a photointerpretation site ........ 25
Figure 6:An example of spatial disaggregation, with image 1 representing administrative boundaries, and image 2 depicting the same recorded census values, after being disaggregated into a gridded format through a weighted surface. ........ 26
Figure 7: Random Forest Model Visual........ 27
Figure 8: Random Forest Model Variable Importance Chart ........ 32
Figure 9: Partial Dependence Plots........ 33
Figure 10: Multi-Variate PDP indicating the relationship between predicted dwellings, road length, and urban/agricultural cell count ........ 34
Figure 11: Predicted Dwellings versus Observed Dwelling Counts (Log-Transformed)........ 36
Figure 12: Predicted Dwelling Distribution for sites with 1-5 Observations, and 6 or more observations, respectively........ 36
Figure 13: Dwelling prediction approaches, scaled up to the entire study area. ........ 39
Figure 14: RF Regression / RF Binary / Nighttime Light (Peace River, Alberta)........ 40
Figure 15: Urban Land Cover Influence on the model ........ 41
Figure 16: Comparing the RF Regression Approach to the Univariate Nighttime Light Approach in Predicting Dwellings........ 42
Figure 17: Study area extent ........ 59
Figure 18: Evaluation 2 Scenario Sites........ 59
Figure 19: Alberta Census Divisions versus Alberta Dissemination Areas. ........ 60
Figure 20: The Dwelling Count Disaggregation Process ........ 65
Figure 21: Global Urban Footprint Dwelling Allocation ........ 68
Figure 22: Nighttime Light Dwelling Allocation ........ 69

Figure 23: Areal Weighted Layer Dwelling Allocation ................................................................ 70
Figure 24: Rural Dwelling Distribution Model Dwelling Allocation........................................... 71
Figure 25:Dwelling Over Allocation (Near Jasper)....................................................................... 73
Figure 26: Event Scenario Sites.................................................................................................... 74
Appendix A.1: HSL Dwelling Allocation Density Plot ……………………………………………82

# CHAPTER ONE: INTRODUCTION

## 1.1 BACKGROUND

### 1.1.1 Introduction

Flooding events are one of the most commonly occurring natural hazards in Canada, taking place across several regions of the country, and causing upwards of hundreds of millions of dollars in damages per event (Davies, 2016; The Canadian Disaster Database, 2018). Decision makers must be able to understand the flood risks that face communities so that they can effectively plan and allocate resources in the case of a future flood events. An important component to understanding flood risk is flood exposure, which describes the number of people and resources, as well as economic and infrastructural assets that could be affected by a flooding event (Hirabayashi et al., 2013). Some possible values that can be used to understand flood exposure include population density and average property value (Elshorbagy et al., 2017). This information can then be processed alongside other key flooding determinants such as flood hazard and flood vulnerability to determine flood risk.

Another way to represent exposure is to measure the number of structures and buildings present in an area impacted by flooding. Specifically, knowing the number of dwellings present in a flooded area can be significant for decision makers, due to its immediate usefulness during a flooding event. With the amounts of displacement and damage to transportation infrastructure, understanding where dwellings are allows decision makers to focus their resources on more heavily impacted areas. This is not only applicable to flooding but can be used to determine and communicate risk for other environmental disasters, such as an industrial accident.

### 1.1.2 Measuring Dwelling Exposure

Accurate dwelling data is essential when trying to understand which parts of a community are exposed to potential hazards. In Canada, this information is collected through the census, taken every five years. These values, including dwelling counts, are publicly accessible, and joined to administrative boundaries at several geographic scales. While this is useful information, it is not always compatible with other spatial data due to differences in feature geometry and georeferencing. Furthermore, these data often lack spatial precision, especially in rural regions. This is due to some census units containing a fairly standard population size regardless of total polygon area. The result is a significant increase in census polygon size as one moves from more dense population centers, where a census unit may be the size of a city block, to more remote and rural areas, where a census unit could potentially be over 100 km$^2$.

Gridded population and settlement layers have been developed over the past several decades, and can provide a more realistic representation of population distribution, especially in regions where censuses are infrequent or unreliable. Human Settlement Layers, or HSLs, are not confined to only using census values, but can use several inputs that represent or are related to human presence. HSLs are normally one of two formats: population-focused, and structure-focused. Population-focused HSLs use census data along with other sets of information to distribute a value (such as population count or density) over a given area. Structure-focused HSLs on the other hand, detect the presence of built-up areas across a given region. This often does not consider population count data, but incorporates additional information such as processed satellite imagery, to generate a gridded surface indicating whether an area is "built up" or "not built up". Both types of HSL are extremely useful for determining human distribution and possible exposure, however there is a gap that is present with regards to capturing the distribution of dwellings. While structure-focused HSLs are designed to capture built up areas, they often do not make distinctions between dwellings and other buildings.

## 1.2 RESEARCH OBJECTIVES

The present research aims to create a dwelling distribution model for rural Alberta. This model can be used to estimate exposure to future flooding events, as well as be used for additional environmental risks such as pipeline failures and industrial accidents that take place over large areas of land. The dwelling distribution model is intended to be region-specific, using publicly accessible spatial data, and uses a random forest framework during its construction. Once created, an external validation of both the rural dwelling distribution model (RDDM), as well as a nighttime light layer will take place against a test data set.

Next, we  compare the RDDM's ability to accurately distribute dwellings over a given region, as well as estimate dwelling exposure through several site-specific event scenarios. In doing so, we aim to see how a region-specific dwelling model compares against less complex and univariate HSLs, as well as HSLs made at a global scale, and whether or not this tool for measuring exposure is a cost-effective approach for decision makers.

## 1.3 CHAPTER OUTLINE

### 1.3.1 Summary

This thesis includes four chapters. Chapter 1 (the introductory chapter) establishes the context for the remainder of the research, as well as explain the purpose of each following chapter. The research is presented in chapters 2 and 3, and lastly, chapter 4 summarizes the findings of chapters 2 and 3 and discusses the implications of their results.

### 1.3.2 Chapter 2

In chapter 2, we create the rural dwelling distribution model (RDDM) by combining several thousand photointerpretation sites with several spatial datasets related to human presence, which are then integrated into a gridded layer covering the study area. Once integrated, a random forest algorithm is used to create a regression model, which will predict dwelling counts across rural Alberta. This model is then evaluated against a test dataset, as well as against a less complex, univariate HSL composed of nighttime light data, to be used as a proxy for determining human presence.

### 1.3.3 Chapter 3

In chapter 3, we evaluate the RDDM's ability to accurate allocate dwellings across a given region, as well as its ability to estimate dwelling exposure in multiple disaster scenarios. Several HSLs are included in this evaluation, including a well-known structure-focused HSL, along with a less complex HSL made only using census data.

### 1.3.4 Chapter 4

In chapter 4, the main findings of both chapters 2 and 3 are discussed, as well as the recommendations for future HSL evaluations.

## REFERENCES

Davies, J. B. (2016). Economic analysis of the costs of flooding. Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques, 41(1–2), 204–219. https://doi.org/10.1080/07011784.2015.1055804

Elshorbagy, A., Bharath, R., Lakhanpal, A., Ceola, S., Montanari, A., & Lindenschmidt, K.-E. (2017). Topography- and nightlight-based national flood risk assessment in Canada. Hydrology and Earth System Sciences, 21(4), 2219–2232. https://doi.org/10.5194/hess-21-2219-2017

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., & Kanae, S. (2013). Global flood risk under climate change. Nature Climate Change, 3(9), 816–821. https://doi.org/10.1038/nclimate1911

The Canadian Disaster Database. (2018, December 21). Available from: https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx

# CHAPTER TWO: CREATING A RURAL DWELLING DISTRIBUTION MODEL FOR THE PROVINCE OF ALBERTA

## 2.1 INTRODUCTION

### 2.1.1 Background

Flooding events are some of the most frequent natural hazards that take place in Canada and are compounded by factors such as increased rainfall and snowmelt (The Canadian Disaster Database, 2018). Floods take place all across the country and have caused damages ranging from over 600 million to approximately 6 billion dollars per event in recent years (Davies, 2016). With the onset of climate change and extreme weather events increasing in numbers and intensity, flooding events are expected to also increase in their frequency and impacts within Canada and across the globe.

As such, decision makers have an incentive to understand the flood risks that their communities face, so that flood mitigating measures can be effectively put into place. A key component of understanding flood risk is flood exposure, which describes the number of individuals, goods, and resources, as well as economic and infrastructural assets that could be affected by a flooding event (Hirabayashi et al., 2013)**.** Examples of values that can be used to understand flood exposure include population density, as well as property value (Elshorbagy et al., 2017). This information can then be processed alongside other key flooding determinants such as flood hazard and flood vulnerability to determine flood risk.

One of the possible ways to calculate exposure is to measure the number of structures (e.g. residences, businesses, and other buildings) that are present in an area damaged by flooding. Knowing the number of dwellings impacted by a flooding event is of great importance to policy makers, as this information has immediate relevance in a natural disaster. For example, during a flooding event, every flooded dwelling would displace people from their homes, and these people would require food, resources, and immediate shelter (Levine, Esnard & Sapat, 2007). Furthermore, flooding events would damage surrounding infrastructure such as roadways and powerlines, and by understanding the distribution of dwellings, policy makers can prioritize reconstruction in more populated regions (Lamond, Booth, Hammond & Proverbs, 2011). Lastly, flooding events can have a long-term economic impact on a region, due to businesses being forced to either temporarily or permanently shut down, directly affecting surrounding workers (Davies, 2016).

This chapter describes the development, implementation, and validation of a method for estimating dwellings as a measure of flood exposure. This process is carried out in rural Alberta, Canada, and incorporates regional spatial data along with dwelling counts collected via photointerpretation. The performance of the method is then compared against a globally collected spatial layer that is often used as a proxy in determining human presence.

### 2.1.2 Measuring Dwelling Exposure

Currently, some private dwelling information is accessible through the Canadian census. Defined as a set of living quarters, dwelling count values are available at the smallest publicly available census unit known as the Dissemination Area, or DA. These polygons cover the entirety of Canada, and while they range greatly in size, they contain an average population of 400 to 700 individuals within them. DAs can be useful for making geographical or temporal comparisons of demographic indicators (Government of Canada, 2016).

In urban and peri-urban locations, DAs are often the size of a city block, but in more sparsely populated areas, DAs can increase in area to over 100 km$^2$. This variation in size can become an issue when using large administrative polygons as representations of flood exposure and integrating them with a rural flooding model. This lack of precision in rural areas results in DA's indicating the number of dwellings in a region spanning several hundred kilometers while giving no further information on the actual dwelling distribution within the DA itself (Wardrop et al. 2018). This is in contrast to the large amounts of high precision elevation and hydrological data that are used in flood model construction.

An alternative approach is to use gridded population and settlement layers. These layers consist of equally sized cells distributed over a given surface, with every cell indicating a value that represents human presence. These values can refer to population density, dwelling counts, or other proxies for human distribution. Unlike census–based polygons that are delineated by administrative boundaries, a settlement layer can be easily integrated with other datasets due to its grid format. An example of a gridded population layer is provided in Figure 1, with a hypothetical set of census units present in the left-most square, and what a resulting gridded population layer may look like on the right. In the case of a flooding event, this alternative approach to indicating population can provide a more accurate representation of exposure, and ultimately  allow for better performing flooding models.

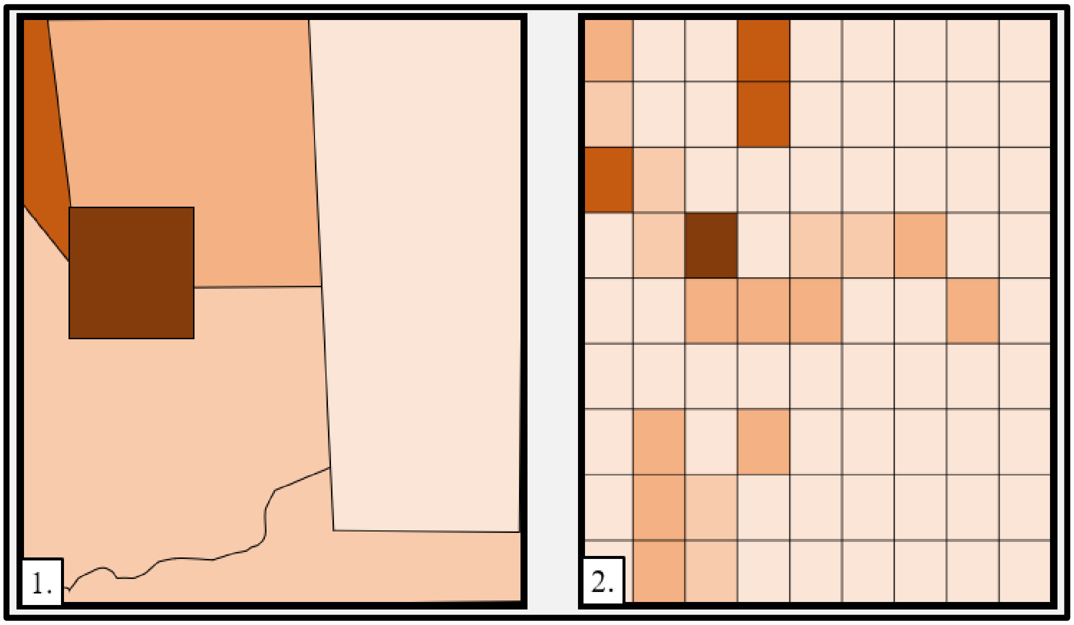


*Figure 1: (1) A hypothetical set of administrative boundaries, indicating a variable related to human distribution (e.g. Population, Dwelling Count). (2) The same area of land, but with the values converted into a gridded format.*

In order to create these products, hence referred to as Human Settlement Layers (HSLs), research groups from around the world have developed methods for integrating different types of spatial data into representations of human population and settlement distribution. HSLs typically have two formats: *population-focused* and *structure-focused*. The population-focused approach uses population count data and distributes it over an area using spatial weights that act as proxies for human presence (CIESIN, 2016). An example of an HSL that uses this technique is

LandScan, which provides an estimate of people that were present in that cell in any given day (Dobson, Bright, Coleman, Durfee, & Worley, 2000). This is completed through dasymetric mapping, which consists of taking several variables related to human presence (such as road density, nighttime light, etc.), using them to create an index representing the likelihood of human presence at a location, and using the index as a filter to transform a set of administrative boundaries into a gridded layer (Dobson et al., 2000). This approach has been adopted in several instances, and the actual LandScan layer has been used as a reference for population layer comparisons in multiple studies (Hall, Stroh, & Payá, 2019; Roy & Blashke, 2014; Stevens, Gaughan, Linard, & Tatem, 2015). Other population-focused HSL's exist, such as the Gridded Population of the World, which simply takes the census counts for an area's smallest administrative boundary, and equally distributes the population count within it, referred to as an "areal-weighting" method (CIESIN, 2016).

The structure-focused approach deals specifically with the presence of built-up areas. This information is collected through imagery data and is then processed into a gridded raster layer (Esch et al., 2017). This method does not usually make use of population data but rather focuses on the captured physical information of an area of interest, and is often configured to have a binary output, indicating that either a structure is present or not (Pesaresi et al., 2013). These types of HSL's are often composed of a smaller number of inputs, often using either optical or radar imagery to base their modelling on, such as the data provided by sources like the Landsat-8 Operational Land Imager, and the Sentinel-2 Multispectral Instrument (Pesaresi et al., 2013). An example of this is the Global Urban Footprint, created by the German Aerospace Center (DLR). A binary surface identifying what is and what is not a built-up area at an approximately twelve-meter resolution, the Global Urban Footprint was created with Radar imagery, collected via DLR-owned satellites (Esch et al., 2017). Once captured, this information was then processed through unsupervised classification, and was passed through several mask layers including features such as waterbodies and road networks (Klotz et al., 2016).

While both these formats can be used to visualize human presence, they differ in which final metric should be used as a proxy for built-up structures. Structurally focused HSLs can generally be produced at a much higher resolution than their population focused counterparts, due to innovations in satellite capture quality, and their overall more straightforward construction (Klotz et al., 2016). Population focused HSLs on the other hand, can provide additional context in their products, through the use of additional variables related to human presence being used to create a weighted surface, which could be used to disaggregate population values (Dobson et al., 2000). This does however, come at the cost of needing to standardize several types of variables, along with the creation of a multi-variate model, which can require a large amount of resources for both computational power, as well as upkeep (Stevens et al., 2015).

When it comes to the construction of HSLs, regardless if it focuses on modelling populations or structures, the fewer variables included in the modelling process, the less maintenance required for keeping it up to date, and the easier it is to use. For example, the Gridded Population of the World uses an areal-weighting method, which equally distributes a population count within an administrative boundary, making it less accurate than LandScan in terms of indicating human presence, but more easily maintained (CIESIN, 2016). Another strategy that researchers have used to address this is to use nighttime light data, which is publicly available and covers the globe. Nighttime light is an effective tool at representing human activities on the earth's surface and is known to have a high correlation with population density (Anderson, Tuttle, Powell, & Sutton, 2010; Liu, Sutton, & Elvidge, 2011). Recently, nighttime light has been used to indicate flood risk across Canada in the form of a composite layer combined with a land classification layer, representing flood exposure (Elshorbagy et al., 2017).

### 2.1.3 Machine Learning and Population Mapping

Today, many HSLs use machine learning frameworks for their model construction. Machine learning is the process of using artificial intelligence to gather information from a dataset, and then use this information to understand the patterns present within the data, with minimal human intervention (Kanevski, Pozdnoukhov, & Timonin, 2008). Some machine learning methods include supervised and unsupervised learning algorithms, with the former using historical data to generate a prediction about unknown events, and the latter using no frame of reference when describing the pattern within the original data (Kanevski, Pozdnoukhov, & Timonin, 2008). Because geospatial data often includes nonlinear relationships, large amounts of variability and outliers, and large numbers of predictors, traditional modelling approaches may be less suitable than machine learning, which is more capable of finding hidden and complex relationships in data (Kanevski, Pozdnoukhov, & Timonin, 2008).

Machine learning has been used to estimate human settlement in previous research. Some approaches use machine learning for image processing.  For example, Weaver et al. (2018) measured the performance of two machine learning approaches, support vector machines, and convolutional neural networks, to identify rural settlements in Afghanistan from high resolution imagery. The work by Hu et al. (2019) builds on this with their mapping of uncounted populations in rural India, where a convolution neural network was used to analyze satellite imagery and predict population densities. Once modelled, LandScan was used as a reference to evaluate against, and this model was seen to outperform the global HSL (Hu et al., 2019).

Another way machine learning has been applied to population mapping is through disaggregating census counts. An example of this is the work by Stevens et al. (2015), who used random forest regression modelling to create a 100m resolution population density map for Vietnam, Cambodia, and Kenya. Through this method, a probability layer was generated and then used as a weighted surface to carry out the disaggregation for each country's respective administrative boundaries. Stevens et al. (2015) noted that the flexible non-parametric nature of this framework allowed them to integrate several different types of spatial variables into the modelling process.

### 2.1.4 The Current Challenge

In urban Canada, census data are probably adequate to estimate exposure; however, there is a need for high-resolution exposure estimation in rural areas. Large amounts of overland flooding has taken place in the provinces of Alberta, Saskatchewan, and Manitoba, provinces with sparse population distributions and large populations of rural-living residents exposed to high flood risk. This historic pattern of flooding combined with populations of people spread out over large areas of land presents a challenge to policy makers that must understand and communicate flood exposure and flood risk in these regions.

Unfortunately, the publicly accessible global HSLs that are currently available may not be adequate to meet the need for high-resolution exposure estimation in rural Canada. For instance, HSLs such as LandScan incorporate a weighted surface used to disaggregate population counts within a region (Dobson et al., 2000). These weights are often generated at the national level, thereby considering large-scale trends in population density while not addressing how these trends differ at the provincial and sub-provincial level. In a country where there are extremely low population counts in areas that simultaneously have high commercial productivity, explanatory variables that are related to commercial activity may have their weights reduced, to minimize overestimation in low population areas (Dobson et al., 2000). A region-specific model may perform better at capturing rural dwelling distribution compared to a model trained at a country-wide level, as the latter would include dense urban areas that would negatively impact performance in more sparsely populated regions.

As well, structure-focused HSL's such as the Global Urban Footprint, while relatively simple to construct, often do not make the distinction between dwellings and any other structure types (Pesaresi et al., 2013). This “all-structures” approach can be problematic when used as a representation of human presence, since commercial and industrial structures would be captured together, providing an imprecise estimate of impact on residents.  Lastly, even though one of the key strengths of some HSL’s are their use of ancillary variables in disaggregating population, these are all collected from either a global or national source (Bhaduri, Bright, Coleman, & Urban, 2007). Having a regional model with regional data and regional model parameterization may better capture smaller scale details than models constructed at a larger scale.

### 2.1.5 Objective

The present research aims to create a dwelling exposure model for rural Alberta. This model could be used in estimating structural exposure to future flooding events in rural Alberta, and may be generalizable to other prairie provinces, rural Canada, and other environmental risks (such as pipeline failures and industrial accidents). This *region-specific* approach will use the same principles as LandScan, incorporating regional ancillary variables related to population and built-up structures to create a weighted surface. A random forest framework will be used to generate an index to disaggregate dwelling counts within census boundaries.

This approach will be compared to nighttime light imagery data, as an alternative tool to disaggregate private dwelling values across rural Alberta. By doing so, we aim to see how a region-specific model performs compared to nighttime light data, and whether this approach is a viable substitute to using a single variable layer. If so, this provides policymakers an alternative approach to communicating flood exposure to their constituents, using publicly accessible information.

## 2.2 METHODS

### 2.2.1 Study Area

The research setting is Alberta, Canada's fourth most populated province, with over 4 million inhabitants. As of 2016, there are over 1.6 million private dwellings in Alberta, with approximately 250,000 of these dwellings present in its rural areas, accounting for just over fifteen percent of the entire province (Government of Canada, 2016). Alberta has a history of flooding, which includes the 2013 Calgary floods, causing approximately $6 billion dollars  in damage (Davies, 2016). Rural Alberta was specifically focused upon due to its lack of high-resolution population data, which could impact the performance of resulting flood exposure models.

The present study defines "rural" as any area not within a designated Town,  Urban Community, or City (hence referred to as TUC), as established by the province of Alberta. Legally, a Town is defined as an area under 1850 $m^2$ containing a population of 1000 or more inhabitants, with a City being defined as being the same area, but containing10,000 or more inhabitants (Province of Alberta, 2020). These administrative boundary layers are made publicly available by the Province of Alberta and were compiled into one merged polygon layer (Province of Alberta, 2016) . This layer was then integrated with a polygon layer indicating rivers and bodies of water within Alberta, and the combined layer was used as a filter, with all areas falling within removed, and the remainder forming the study area. This is shown in Figure 2:

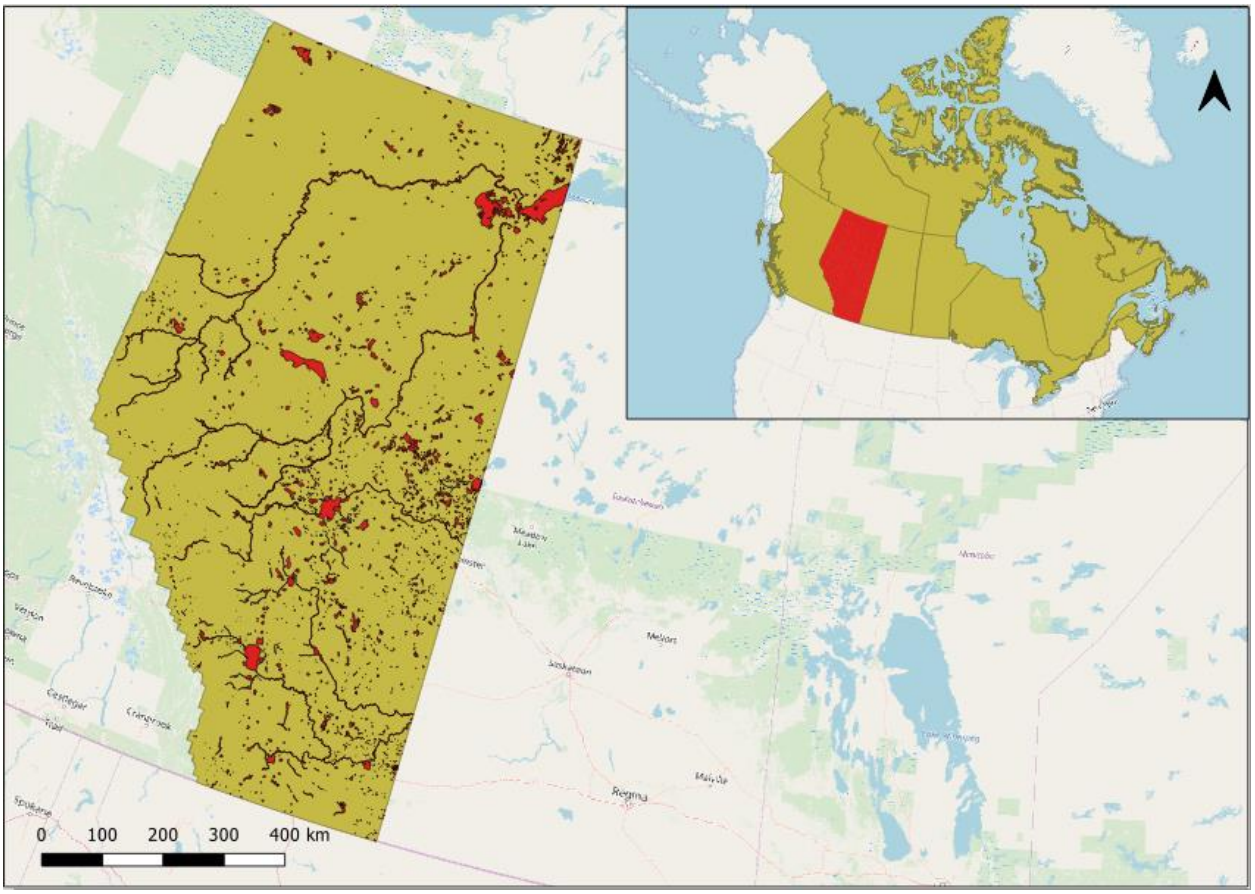


*Figure 2: The extent of the study area (khaki) with areas in red indicating removed sections*

### 2.2.2 Grid layer:

We based the uniform grid layer used in this study on the Dominion Land Survey (DLS). The DLS was a system used to partition Western Canada into square, one-mile sections for agricultural purposes (Robert & McKercher, 1992). The smallest grid unit of the Dominion Land Survey, the quarter section, was selected as the base grid size, with each cell side being a half-mile (or approximately 805 meters) in length. Using a quarter section grid provided by the province of Alberta as an initial layer, all quarter section cells that overlapped with a TUC or waterbody were removed from the study area, with the remainder being used in the modelling process, consisting of over 955,000 cells (Figure 3).

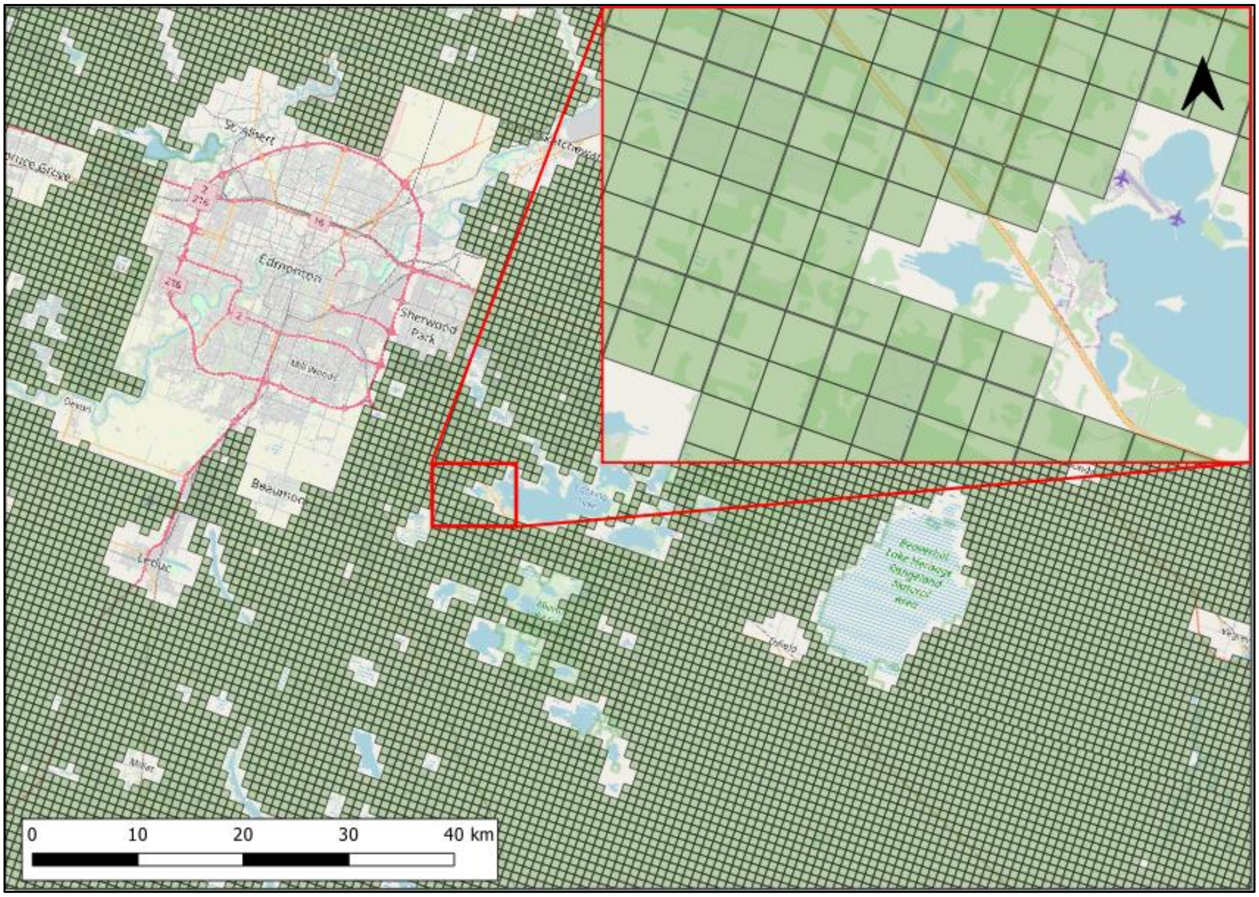


*Figure 3: The Dominion Land Survey, configured to become our gridded study area (Central Alberta) (With grid sections overlapping with a TUC or water body removed)*

### 2.2.3 Data

Several publicly accessible spatial datasets related to human presence are used in this study. These datasets ranged from raster surfaces to polygon and polyline values, and as such, data processing was required prior to directly integrating these layers to the study area grid. This was carried out in R, and the integrated values would then be used to predict dwelling counts later on in the modelling stages.

#### 2.2.3.1 Raster Data

*Landcover data*

We use a 2015 Canada-wide land cover map published by Natural Resources Canada. Publicly available in a 30-meter resolution layer, and based off of processed LandSat 8 imagery, this dataset consists of several classes including agricultural land, urban land, and additional ecozones such as forests and wetlands (Government of Canada, 2019). Prior to being joined to the quarter sections, the land cover data was resampled up to a 50 meter resolution, and then was separated into two separate binary layers, with the first layer indicating a value of one where "urban" cells were present, and zero otherwise, and the second layer repeating this process for "agricultural" coded cells. Only the urban and agricultural cells were retained as they were seen as most significant to determining rural dwelling presence. The total number of urban and agricultural cells present in each quarter section was then counted and joined to its respective grid. Proximity values (in meters) indicating the distance from the center of each grid to the nearest urban or agricultural cell were also included.

*Nighttime light*

Nighttime light data was collected alongside land cover to determine dwelling counts and was produced by the Earth Observations Group at the National Oceanic and Atmospheric Administration (NOAA). This specific dataset was created using the Visible Infrared Imaging Radiometer Suite, known as VIIRS, and contained the global temporal average for the year 2016, at a resolution of 15 arc-seconds, approximately 463 meters at the equator (National Centers for Environmental Information, 2016). For the province of Alberta, this results in a resolution of approximately 272 meters. The average nighttime light radiance value was calculated per quarter section and was spatially joined to its respective grid.

*Elevation*

The final raster layers that were joined to the grid layer included a 100m resolution elevation layer encompassing the province of Alberta, and a slope layer produced from this elevation layer, indicating the steepness of an area at the same resolution (Province of Alberta, 2017b). Both elevation and slope have been used in several models depicting human presence, such as LandScan, which notes an inverse relationship between slope and population density (Dobson et al., 2000). As the Canadian Rockies are a key geological feature spanning across the western half of Alberta, understanding the relationship between elevation and slope and rural dwelling count is essential.

#### 2.2.3.2 Vector Data

*Hydrological and Transportation Data*

A vector line dataset indicating streamflow throughout Alberta was integrated with the grid layer by measuring the total stream length present within each cell's boundary, with the total value in meters joined to each respective cell (Province of Alberta, 2017a). This dataset was originally created to support catchment area delineation and was retrieved from Alberta's natural resources data archive. Similarly, the total length of roadways present in each cell was recorded and integrated into the grid layer, as transportation infrastructure has been a widely used indicator in population modelling (Dobson et al., 2000; Esch et al., 2018; Wardrop et al., 2018). Building on this, rail lines were also measured within each grid cell. Both the road and train data was collected from the Canadian federal natural resources data repository.

*Fossil Fuel Industry Data*

Several variables that represent industrial build-up and commercial presence were also integrated with grid layer. With Alberta being Canada's largest producer of crude oil, as well as possessing a large number of mines, refineries, and coal plants throughout the province, the sheer size of this industry results in large amounts of land being classified as built-up or urban, as well as producing large amounts of nighttime light (Government of Canada, 2020). However, this occurs without any actual dwelling existing on these properties. These variables include an oil and gas pipeline layer made available by the Alberta Energy Regulator, whose total length per cell was recorded and integrated into the grid layer, as well as an oil and gas well-pad layer, with the total number of well-pads per quarter section joined to each respective cell (Alberta Energy Regulator, 2018). Lastly, the locations of all the refineries, mines, gas fields, natural gas storage facilities, and power plants over 100 megawatts were included in the form of point coordinates.

This layer was retrieved from the North American Cooperation on Energy Information dataset (Government of Canada, 2017b).

*Indigenous and Crown Land Data*

The administrative boundaries of the Indigenous Lands of Canada was used to demarcate the boundaries of Indian Reserves, Land Claim Settlement Land, and Indian Lands (Government of Canada, 2017a). As these areas span large regions of space in rural Alberta, it was deemed appropriate to include them in the modelling process to aid in predicting dwelling count. Crown land boundaries, which refers to property that is owned by either the federal or provincial government, was incorporated using data that are publicly accessible through the province of Alberta's data repository, and was included in the modelling process due to its association with low census counts (Stevens et al., 2015).

With regards to the extraction industry layers, as well as the Indigenous Land  and Crown Land layers, proximity values were generated for each grid cell in the study area, indicating the distance from its centroid to the nearest industrial site, Crown Land, and Indigenous Land, respectively. Prior to each dataset being processed and integrated into the grid layer, they were first projected to the same coordinate system used throughout this research, "NAD83 / Alberta 10-TM” (EPSG: 3400). This projection is commonly used by the government of Alberta and was the format the majority of the province-specific data was already published in, so all additional data were projected to maintain consistency (Province of Alberta, 2010).

Next, a term comprised of the number of urban cells in a grid multiplied by the number of agricultural cells was generated. This was done to capture any relationship that may exist when both are present, such as when a dwelling exists on a large plot of farmland, versus either of those two types of land classes existing alone. Lastly, quadratic versions of both the urban cell count variable and the nighttime light variable were created. All data preprocessing and analysis was carried out in QGIS, ArcGIS, and R, with the data integration taking place exclusively in R. A detailed summary table on the variables is included in Table 1.

*Table 1: Explanatory Variables*

| Spatial Data Type: | Name: | Description: | Source: |
|---|---|---|---|
| **Binary Raster** | **Urban Binary Layer** | 2015 - Canada Land Cover Layer:<br>• Recalculated to return either "Urban" or "Not Urban" land<br>• Originally 30m resolution (resampled to 50m)<br>• Number of cells per grid counted<br>• Distance to the nearest urban cell recorded from grid center | Government of Canada |
| | **Agricultural Binary Layer** | 2015 - Canada Land Cover Layer:<br>• Recalculated to return either "Agricultural" or "Not Agricultural" land<br>• Originally 30m resolution (resampled to 50m)<br>• Number of cells per grid counted<br>• Distance to the nearest agricultural cell recorded from grid center | |
| **Continuous Raster** | **Elevation** | 2018 - Elevation Layer for the province of Alberta<br>• 100 m resolution DEM | Province of Alberta |
| | **Slope** | Generated from the 2018 Elevation Layer for the province of Alberta<br>• 100 m resolution slope layer | |
| | **Nighttime Light** | 2016 - Nighttime light layer<br>• Specifically used  the following:<br>(VIIRS Cloud Mask - Outlier Removed - Nighttime Lights) contains the "vcm-orm" average, with background (non- | National Centers for Environmental Information (NCEI) \| National Oceanic and Atmospheric Administration (NOAA) |

| | | | |
|---|---|---|---|
| | | lights) set to zero.• Selected and clipped out the specific region being studied | |
| **Vector Polygon** | **Towns / Urban Centers / Cities (TUC)** | Used in the creation of the study area:<br>• Distance from quarter section centroid to closest TUC | Province of Alberta |
| | **Crown Reservation Land** | 2017 - Crown Reservation Land (registered as publicly owned land)<br>• Distance from quarter section center to closest Crown Land polygon | |
| | **Parks and Protected Areas** | (Revised 2016) Provincial parks and environmentally protected regions within Alberta<br>• Distance from quarter section center to closest polygon | |
| | **Waterbodies** | 2019 - Waterbody polygons within Alberta<br>• Used in the creation of the study area | |
| | **Indigenous Land** | (Revised 2019) The Indigenous Lands layer depicts the administrative boundaries (extent) of lands where the title has been vested in specific Aboriginal Groups of Canada or lands which were set aside for their exclusive benefit.<br>• Distance from quarter section center to closest polygon | Government of Canada |

| | | | |
|---|---|---|---|
| **Vector Polylines** | **Roads** | (Revised 2019) - Canadian National Road Network<br>• Total number of lines within each quarter section | Government of Canada |
| | **Train Tracks** | (Revised 2018) - Canadian National Rail Network<br>• Total number of lines within each quarter section calculated for each unit | |
| | **Pipelines** | 2018 - Pipeline Layer, containing all Oil and Gas pipelines approved by the Alberta Energy Regulator<br>• Total number of lines within each quarter section | Alberta Energy Regulator |
| | **Stream and Water Network Lines** | 2018 - Hydrographic Network of Alberta<br>• Total number of lines within each quarter section calculated for each unit | Province of Alberta |
| **Vector Point Files** | **Producing Mines / Oil & Gas Fields** | 2019 - Principal Mineral Areas, Producing Mines, and Oil and Gas Fields<br>• Distance from quarter section center to closest point | Government of Canada |
| | **Refineries** | (Revised 2018) Refineries - North American Cooperation on Energy Information<br>• Distance from quarter section center to closest point | |
| | **Natural Gas Underground Storage** | 2017 - Indicating underground facilities used for storing natural gas - North American Cooperation on Energy Information<br>• Distance from quarter section center to closest point | |
| | **Power Plants** | 2017 - All Power Plants with an installed capacity of 100 megawatts or more<br>• Distance from quarter | |

| | | | |
|---|---|---|---|
| | | section center to closest point | |
| | **Well pads** | 2016 - Oil and Gas Well pads in use<br>• Total number of pads within each quarter section counted and joined to each spatial unit | Province of Alberta |

### 2.2.4 Data Collection via Photointerpretation:

Next, approximately 12,500 grid cells were randomly selected from the study area, and their respective dwelling counts were recorded via photointerpretation. These cells are spread throughout the province of Alberta, capturing a variety of different environments, visualized in Figure 4:

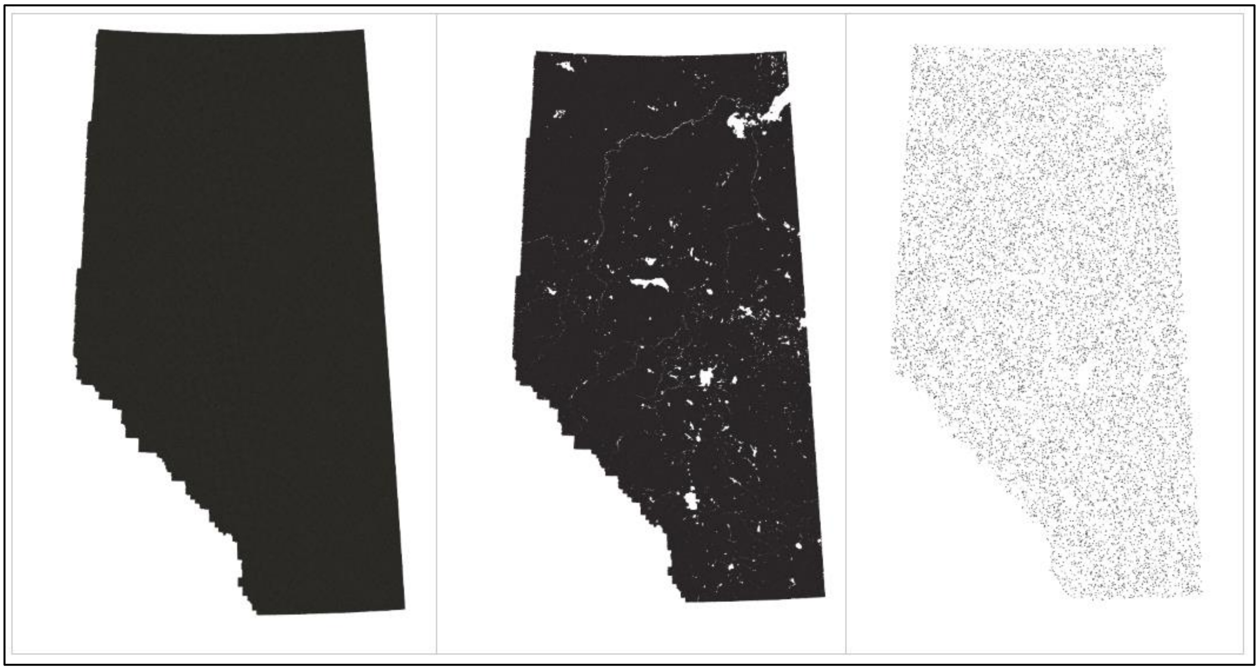


*Figure 4: From left to right - The entire province, the study area (No overlap with Towns / Urban Centers / Cities and waterbodies), and the randomly selected cells.*

The data collection was carried out using Google Earth Imagery, with the image classification process working as follows. For every site, the entirety of the quarter section was visually inspected, with the number of observed residential dwellings recorded. Cells that had no structures present were assigned a ‘0’ value. For cells that included several buildings, such as a commercial storage facility, non-residential buildings were not counted during data collection, with the aim being to only record the number of dwellings physically present. Ignored structures included barns, silos, and other non-dwellings.

Several visual cues were used to distinguish between residential and non-residential buildings.  These include the structure's size, shape, colour, relative location in the cell, proximity to a main road, and immediate vegetation surrounding the structure, such as trees or shrubs. An example of this is provided in Figure 5. In this example, a large plot of land being primarily used for agriculture is featured, with several structures present in its bottom-right corner. Examining these buildings more closely, the previously mentioned visual cues are present with regards to one structure, that has vegetation immediately surrounding it. This structure also has a distinct colour, in comparison to the surrounding plain-metal buildings. In this scenario, the entire cell would have a value of "one" recorded. Once completed, recorded dwelling counts are then joined with their respective grid cell, alongside the previously mentioned spatial layers.

This approach has been used at a larger scale during the construction of the World Settlement Footprint, where the creators of the Global Urban Footprint, in partnership with Google, had 900,000 sites validated using crowdsourced photointerpretation (Marconcini, 2019). Using this data source maintained the overall research aims of committing to the use of open source and publicly accessible data, while also leveraging the very high-resolution imagery provided by Google Earth's imagery collection. The imagery data used for dwelling count collection was from between 2008 and 2018, and is composed of several imagery sources, varying in resolutions from medium to very high, including LandSat 8, Digital Globe's Worldview Series satellites, as well as collected airborne data (Marconcini, 2019).

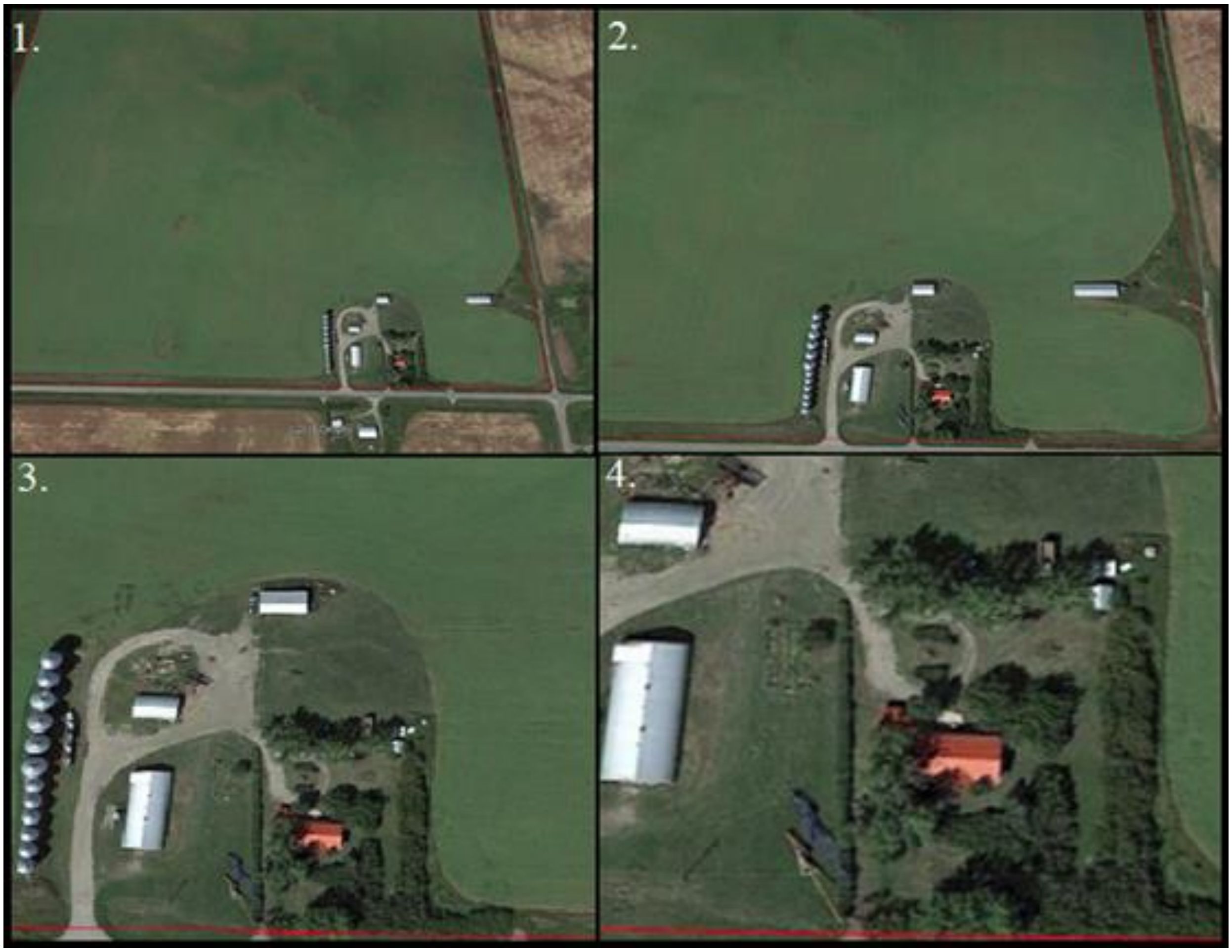


*Figure 5: An example of a photointerpretation site (with an increased magnification from photo 1 to 4).*

### 2.2.5 Modelling:

After the ancillary variables were joined to the grid layer, and the dwelling count data was collected via photointerpretation, our aim was to create a model to predict the number of dwellings present within a quarter section. This model would be trained using 70% of the photointerpretation sites, with the remaining 30% being used for evaluation. Once created, the model would then be applied to the entire study area, using the spatially joined explanatory variables to create a dwelling count prediction for each cell. The resulting layer would not be directly used to represent dwelling counts, but would act as a weighted surface, which would then disaggregate census-recorded dwelling count values from their original Dissemination Areas into a gridded layer. This is visualized in Figure 6:

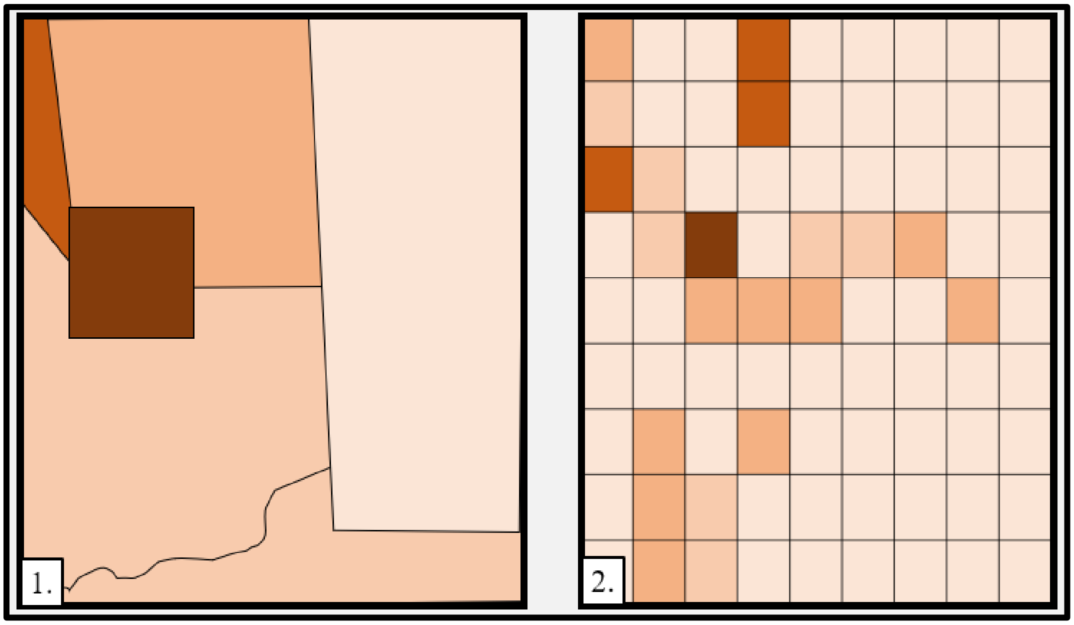


*Figure 6:An example of spatial disaggregation, with image 1 representing administrative boundaries, and image 2 depicting the same recorded census values, after being disaggregated into a gridded format through a weighted surface.*

We utilized a random forest algorithm to predict dwelling count values, with the recorded dwelling counts being the dependent variable, and the ancillary spatial layers being explanatory variables. The random forest algorithm itself is an ensemble method, due to it aggregating the results of several decision trees and determining a final value from the average of all the generated trees.

Each of the decision trees that make up the random forest are constructed as follows. The training dataset is continuously split from a starting "root" node in a way that results in the largest drop of entropy by the terminal (or final) node. Entropy in this instance refers to the amount of variability that is present in the data, with a high entropy value indicating a high amount of variability. These splits are determined by setting true/false parameters, such as "is X greater than 0.01", and having each answer be its own split, or branch, in the dataset. While relatively easy to interpret, decision trees often have issues with overfitting, which is resolved by using the Random Forest approach (Brieman, 2001).

Rather than rely on a single tree, this technique uses bootstrap aggregating, or bagging, to randomly select data from the training set with replacement, to independently train each tree that is created. By calculating the results of several hundred decision trees, we are able to greatly decrease the amount of variance in the final model, due to each tree being independently created. Since we are carrying out regression, the output of each tree is averaged together, and this value is our predicted result (Figure 7).

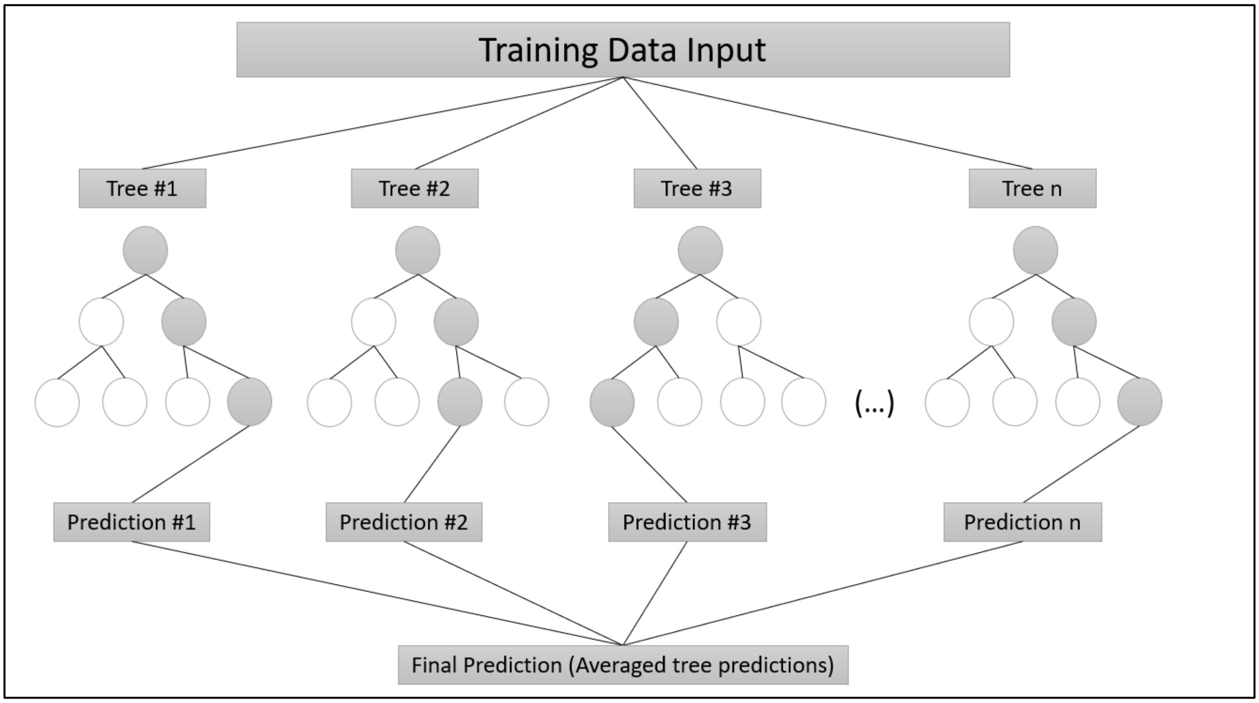


*Figure 7: Random Forest Model Visual*

Model creation and fitting was carried out in R, using the randomForest package (Liaw & Wiener, 2002). In the construction of the model itself, the following three parameters are of great importance: the number of decision trees grown, the amount of observations present in each decision tree's terminating node, and the number of explanatory variables selected for each tree's growth. For the purposes of this study, the default settings used for the number of trees in the model (500) were sufficient, and the recommended minimum number of terminal nodes was used, which is 5 nodes for random forest regression. Once the initial regression model was trained, the tuneRF function was used to optimize the number of explanatory variables that were randomly selected for each tree's growth, which resulted in 24 variables being used for each tree's construction, compared to the originally set 8 variables.

Once the random forest model is created, variable importance is determined for every explanatory variable by looking at the total decrease in node impurity that occurs when a branch splits. Node impurity refers to how well a decision tree splits the data. For regression, this is done by measuring the difference in the tree's residual sum of squared errors before and after splitting on a specific variable. The sum of each split on this variable is calculated for each tree and then averaged across the entire collection of decision trees, with a high value indicating a higher variable importance, and vice versa.

Next, a partial dependence plot (PDP) is generated to illustrate the effects that each predictor has on the outcome of the random forest model (Friedman, 2001). PDPs are a tool used to interpret the results of machine learning algorithms, by rebuilding the random forest model and averaging every predictor except one and telling us for any of the given value of explanatory variable, what the effect on the prediction is.

Alongside the multivariate approach using the random forest algorithm, a univariate dwelling model was also created to predict dwelling counts based on average nighttime light values. This alternative was generated to act as a benchmark to compare with  the random forest approach. To make the comparison, all grid cells where average nighttime light values were greater than "zero", were assigned a value of "one", indicating the presence or absence of light. Next, the same process was carried out with the model output, with all predicted values greater than 0 changed to 1, and the remainder staying as 0. While this would lose some complexity in the final model performance, this binary version of the random forest model allows us to effectively compare these two approaches.

### 2.2.6 Validation and Study Area Application:

The validation of the model is separated into two sections, starting with analyzing the output of the random forest regression model when compared to the testing set, which will be referred to as the "internal validation". Next, the predictive performance of the converted binary random forest approach was compared to the univariate nighttime light approach, with both the nighttime light values. This second category will be referred to as the "external comparison".

Error measurements were calculated for both the internal validation and external comparison, including root mean square error (RMSE) as well as confusion matrices. The following indicators are used in determining classification performance:

**Accuracy -** The total number of correct predictions divided by the total number of predictions made by the model

**Sensitivity** - (True Positive Rate) - Number of correct positive predictions divided by total number of true positives

**Specificity** - (True Negative Rate) - Number of correct negative predictions divided by the total number of true negatives

Both the internal validation and external comparison results are separated into three distinct spatial categories, ordered by proximity to the nearest TUC. These categories are "Urban", including sites that are within 25 kilometers of a TUC, "Peri-Urban", including sites that are within 25 and 50 kilometers away from a TUC, and "Rural", which includes any site greater than 50 kilometers away from a TUC. By completing a stratified evaluation, we will be able to evaluate model performance as one moves farther away from a Town, Urban center, or City.

Lastly, the random forest regression model is applied to the entirety of the study area, to predict a dwelling count value for the over 955,000 cells. This layer is outputted in the form  of a raster grid, with a resolution of 830 meters, using the same projection as the explanatory variables "NAD83 / Alberta 10-TM (EPSG: 3400)". The same resolution and projection is used for the nighttime light approach as well. Note that while we are predicting these values, the resulting dwelling counts will not be directly used in terms of measuring exposure. Rather, these outputs will used as a weighting layer, to disaggregate census dwelling count values in the future.

## 2.3 RESULTS

### 2.3.1 Photointerpretation of Quarter Sections:

With regards to the 12,357 randomly selected grid sites, the vast majority of the cells (~92%) had no structures present. The majority of the cells that did have dwellings present (~8%), had approximately 1 to 6 dwellings recorded within them. In this 8%, there were a few outliers present, including cells with over 20 dwellings counted within them. Furthermore, two of these cells contained what were essentially suburban neighborhoods and had over 200 recorded dwellings within them. Table 2 presents a distribution table of the entire collected dataset.

*Table 2: Recorded Dwelling Frequencies*

| Number of Dwellings Observed | Frequency |
|---|---|
| **No Dwellings Present (0)** | 11352 |
| **1-5** | 946 |
| **6-10** | 27 |
| **11-100** | 30 |
| **Greater than 100 Dwellings** | 2 |

### 2.3.2 Descriptive Statistics of Ancillary Spatial Data:

A total of 24 indicators were included in the model construction (Table 1). Of these variables, summary statistics have been generated for the predictors that have been seen in previous research to have a strong relationship with human presence (Table 3). With regards to every variable included other than 'Distance to the Nearest TUC', we see a distribution similar to that of "Recorded Dwellings", with each indicator skewing to the right, with the median recorded value of each indicator being less than the average.

*Table 3: Descriptive Statistics of Key Predictors*

| | Total Urban Cells | Average Nighttime Light Value | Total Road Length (M) | Distance to Nearest TUC (Km) | Total Agricultural Cells | Urban/Agricultural Interaction |
|---|---|---|---|---|---|---|
| **Mean** | 1.07 | 0.18 | 87.69 | 90.15 | 54.79 | 61.92 |
| **Median** | 0.00 | 0.00 | 0.00 | 52.38 | 0.00 | 0.00 |
| **Mode** | 0.00 | 0.00 | 0.00 | N/A | 0.00 | 0.00 |
| **Standard Deviation** | 10.45 | 3.38 | 313.22 | 91.15 | 95.14 | 440.30 |
| **Minimum** | 0.00 | 0.00 | 0.00 | 1.58 | 0.00 | 0.00 |
| **Maximum** | 260.00 | 286.00 | 8370.00 | 432.81 | 279.00 | 13420.00 |

### 2.3.3 Model Output:

Once the multivariate random forest model was completed, a variable importance chart was generated to allow for a deeper understanding of the impact each variable had in the model's performance (Figure 8). Out of the 24 variables used in the modeling, road length was measured to be of highest importance by a considerable margin, with the urban/agriculture interaction term, and the squared variables being the next most important. All other data were all extremely low scoring relative to road length, none had a value below zero and were kept in the final model.

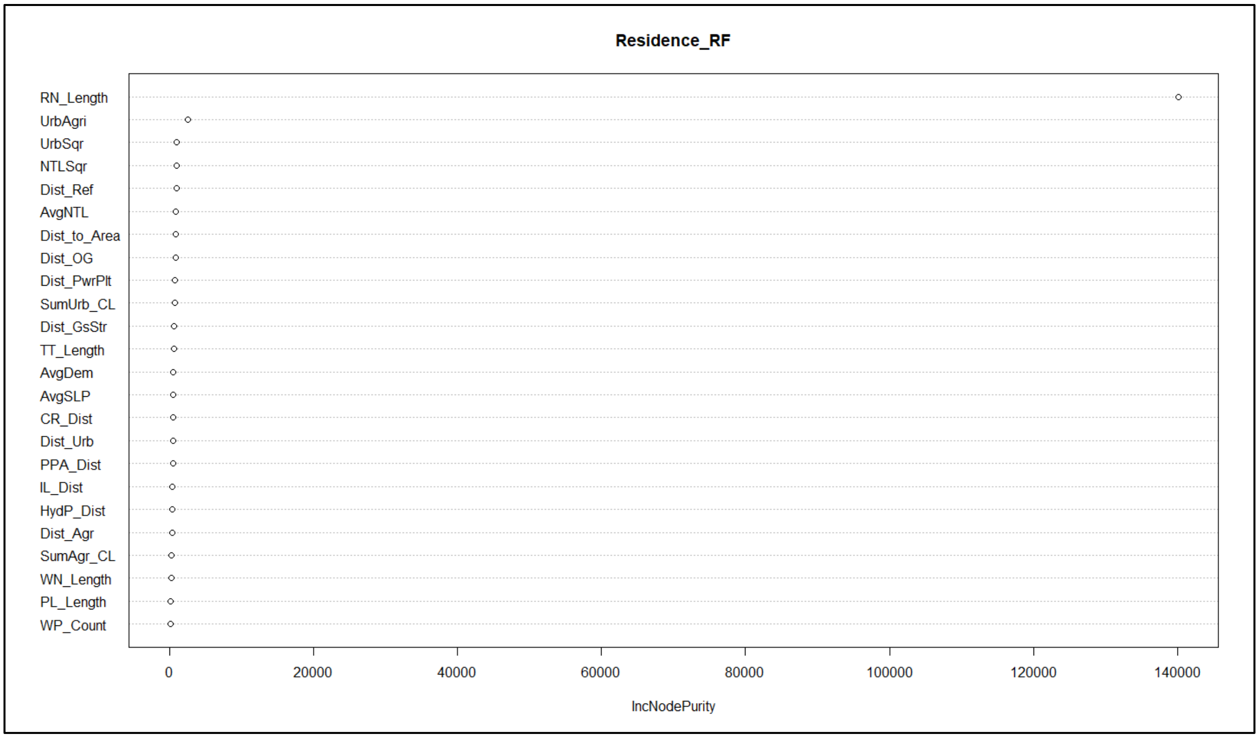


*Figure 8: Random Forest Model Variable Importance Chart*

Next, figure 9 visualizes the relationships between predicted dwellings and the key variables used to predict dwellings, such as measured road length. PDP's were generated for the following variables related to dwelling presence: Road Length, Nighttime Light, Urban/Agricultural Cell Presence, and Distance to the Nearest Town, Urban Center, City (Figure 9).

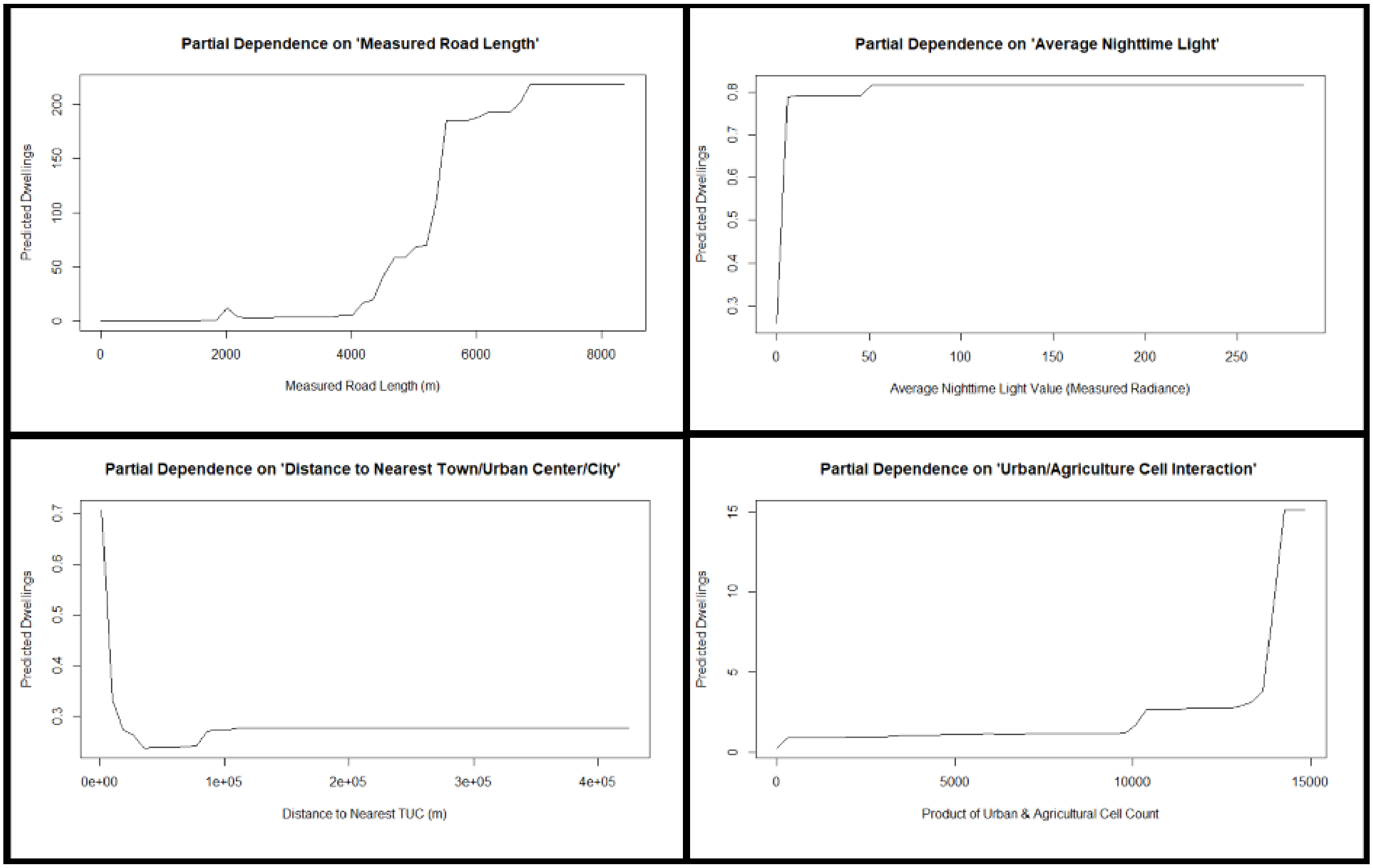


*Figure 9: Partial Dependence Plots for (Clockwise starting from the upper left image): Measured Road Length, Recorded Nighttime Light, Urban/Agricultural Cell Interaction, Distance to Nearest TUC*

For road length there is a slight increase in predicted dwellings at about 2000 meters of recorded road within a cell, followed by an exponential jump after approximately 6 kilometers of roads are recorded within a cell. This is not the case for nighttime light measurements, as there is an immediate jump in the proportion of predicted dwellings once any light is observed. The opposite is seen for the proximity to the nearest TUC, with an immediate drop after about 2 kilometers. Note however that for the 2nd and 3rd partial dependence plots, there is a much smaller increase in predicted dwelling, suggesting that both nighttime light and distance to the nearest TUC do not have a large impact on the predicted output, once all other variables are averaged, which is corroborated by the variable importance chart (Figure 8). For urban and agricultural cell interaction (the product of the number of urban and agricultural cells within each quarter section), there is an almost stepwise increase in predicted dwellings, with a large jump at approximately fourteen thousand, suggesting that for quarter sections with a combination of both agricultural cells and urban cells, the model predicts an increase in the number of dwellings present.

Lastly, we are able to visualize the partial dependence between two explanatory variables and their relationship with target variable. Road length and urban/agricultural cell interaction were measured simultaneously, due to both being key dwelling indicators, and having a similar PDP output. When graphed together, the multi-variate PDP suggests that measured road length is considerably more influential in determining dwelling count, as there is not any notable vertical variation in colour, further supporting the importance of the road length variable (Figure 10).

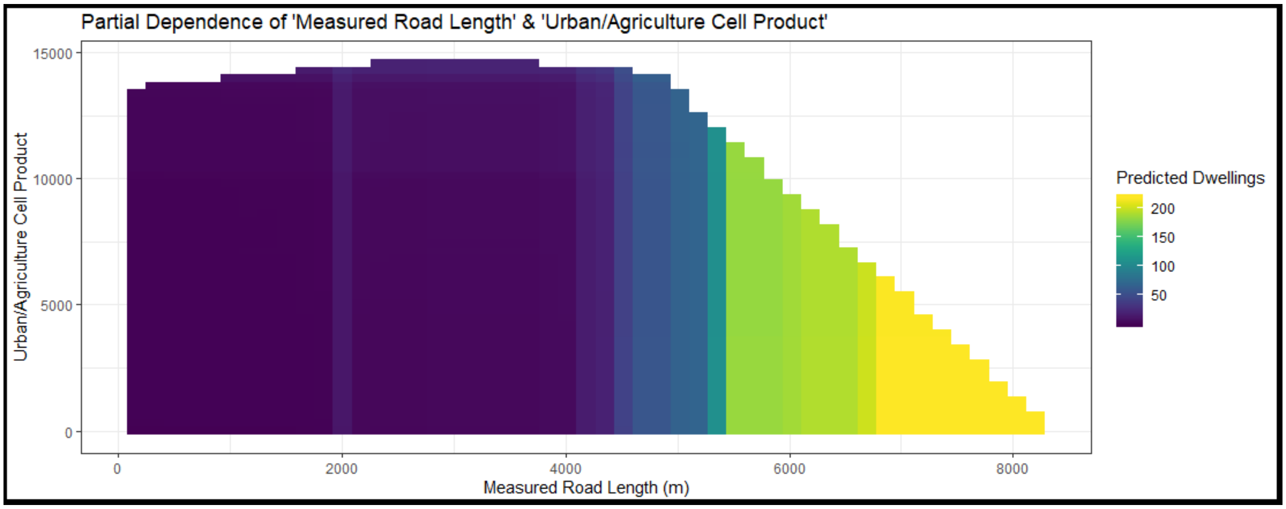


*Figure 10: Multi-Variate PDP indicating the relationship between predicted dwellings, road length, and urban/agricultural cell count*

When evaluated against the test set, the multivariate model performs similarly at the both Province-wide level, and in each stratified region (Table 4). In terms of separating the results by proximity to a TUC, the first stratified region (less than 25 km to a TUC) presents a higher amount of error, decreasing from 2.8 to 0.12 as we move to the furthest region. This trend is most likely due to variability in dwelling counts, which would greatly increase as one gets closer to a town or city, while areas greater than 50 km away from a TUC would often have next to no dwellings present throughout.

*Table 4: Regression Model Results*

| **Regression Model Error - Looking at RMSE** | | |
|---|---|---|
| **Region:** | **Proximity to TUC** | **RMSE** |
| **Entire Province** | **Overall** | **1.402919** |
| **Urban** | **Strat 1 - (0-25km)** | **2.834739** |
| **Peri-Urban** | **Strat 2 - (25-50km)** | **1.412402** |
| **Rural** | **Strat 3 - (>50km)** | **0.1194238** |

Figure 11 presents the observed dwelling counts compared to models predicted values, with a log transformation. Here we see that for the sites with 1 to 3 dwellings present, a predicted value that was somewhat close to the original value was generated, although these were more often underpredicted (Figure 11). This trend of underpredicting dwellings continues for the sites with  4 to 7 recorded dwellings as well. In the few sites with higher numbers of dwellings, the multivariate model still does not perform as very well, with the majority of them still being underpredicted.

However, while the exact counts are not predicted accurately, the overall distribution of the recorded dwellings is captured. This is visualized in Figure 11, where we have the predicted dwelling counts versus the observed value, for quarter sections where at least one dwelling was recorded. To further understand the distribution in the predicted values, Figure 12 presents histograms that were generated specifically for sites that had 1 to 5 dwellings recorded, and 6 or more dwellings recorded, with tick marks along the horizontal axis representing how many instances of each prediction were recorded. In both histograms, we see the underpredicting present within the model, with the upper histogram showing the bulk of the predictions being less than two predicted dwellings, versus being spread out between 1 and 5.

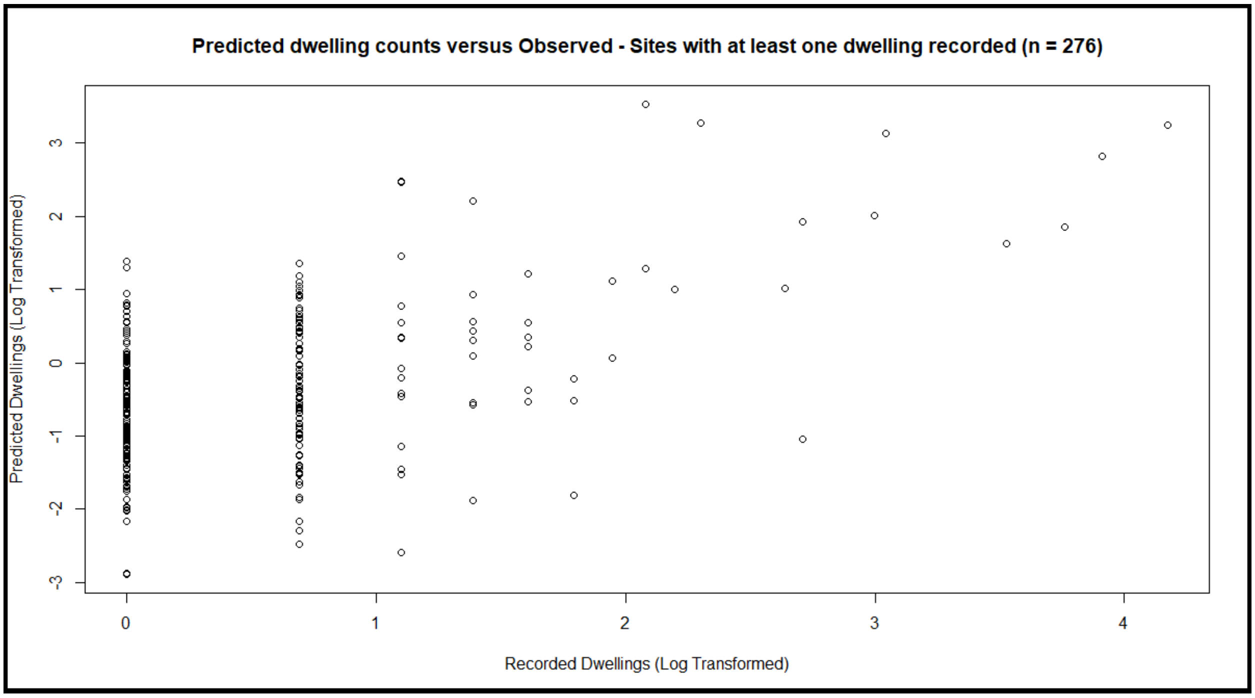


*Figure 11: Predicted Dwellings versus Observed for sites with at least one dwelling recorded (Log-Transformed)*

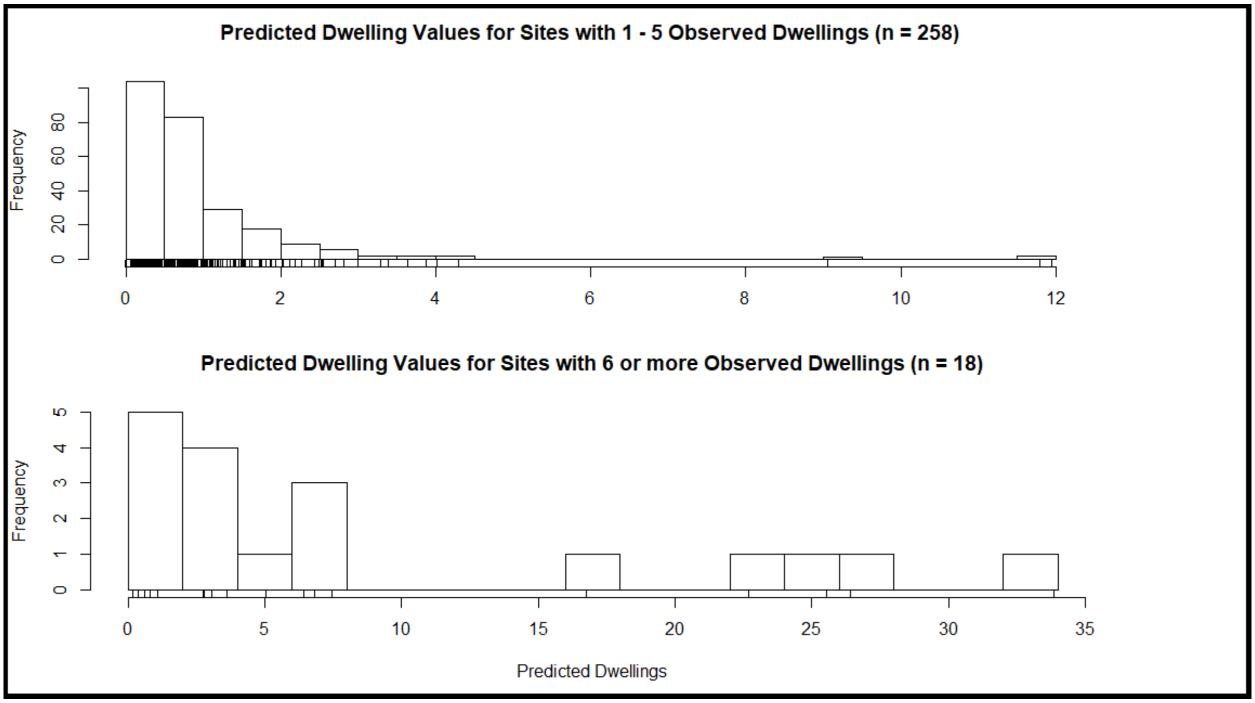


*Figure 12: Predicted Dwelling Distribution for sites with 1-5 Observations, and 6 or more observations, respectively*

Next, we compare the binary random forest model to the univariate nighttime light approach using confusion matrices (Tables 5 & 6). In terms of accuracy, the RF approach outperformed the NTL approach in each distance band, with the widest margin taking place in the areas closest to a TUC. As well, both methods share the same trend of increasing in accuracy the further away from a TUC they go, similar to the regression results. In terms of specificity, the RF approach outperformed the nighttime light approach overall, but in each of the three distance bands, the nighttime light approach was more successful in correctly predicting "no dwelling" sites. Lastly, the random forest model was the better method in terms of sensitivity, outperforming the nighttime light method in correctly predicting "dwelling present" sites when compared to the total number of sites with actual recorded dwellings.

*Table 5: Accuracy Table from Binary Approaches*

| **Binary Model Comparisons - Accuracy** | | | |
|---|---|---|---|
| **Region** | **Distance to TUC** | **RF Accuracy** | **NTL Accuracy** |
| **Entire Study Area** | **Overall** | **0.9356** | **0.8935** |
| **Urban** | **Strat 1 - (0-25km)** | **0.7723** | **0.699** |
| **Peri-Urban** | **Strat 2 - (25-50km)** | **0.92** | **0.8546** |
| **Rural** | **Strat 3 - (>50km)** | **0.9964** | **0.9771** |

*Table 6: Specificity & Sensitivity Table from Binary Approaches*

| **Binary Model Comparisons - Specificity and Sensitivity** | | | | | |
|---|---|---|---|---|---|
| **Region** | **Distance to TUC** | **RF Specificity** | **RF Sensitivity** | **NTL Specificity** | **NTL Sensitivity** |
| **Entire Study Area** | **Overall** | **0.6667** | **0.944** | **0.4964** | **0.9257** |
| **Urban** | **Strat 1 - (0-25km)** | **0.3721** | **0.9232** | **0.564** | **0.75** |
| **Peri-Urban** | **Strat 2 - (25-50km)** | **0.1443** | **0.966** | **0.3918** | **0.9** |
| **Rural** | **Strat 3 - (>50km)** | **0** | **1** | **0.28571** | **0.97957** |

### 2.3.4 Scaling the Models to the Study Area

After evaluation, all three approaches were scaled up to the entire study area. For the random forest models, this was done using the same explanatory variables that were used in the original model training, resulting in a continuous and binary output for the entire study area. For the nighttime light approach, the average nighttime light values for each quarter section were recorded, and then was converted to either show "light recorded" or "no light recorded". All three approaches are shown side by side in Figure 13, with the continuous output on the farthest left, followed by the binary output, and the nighttime light layer, respectively. For both RF approaches, we see high number of predicted dwellings bordering the largest cities in the province, Edmonton, and Calgary, as well as throughout the Edmonton-Calgary corridor. As well, we observe a fairly consistent spread of dwellings following major roadways, outwardly spreading away from any TUCs. Outside this region, there are sparse pockets of dwellings seen throughout Alberta at this scale, with relatively large amounts of predicted dwellings seen in the central west parts of the province, near the Grand Prairie region, as well as south near the US/Canada border (Figure 13 – A&B). Looking at a more sparsely population area of the study area, we can more easily see how each of these approaches differ as a potential final product (Figure 14). Overlooking the Peace River area of North-West Alberta, we can see an overall increase in dwelling count values as we go from the RF regression model to the RF binary model, to the final nighttime light approach.

Of the two RF outputs, the binary approach seems to predict an overall larger distribution of dwellings, although these seem to be all concentrated around the population hubs of the province. Meanwhile, the RF regression method seems to have less dwellings counted within it but does however seem to capture dwellings in relatively more rural regions, such as North-West of Edmonton, and near Fort McMurray in the northeast.

The nighttime light output follows a slightly similar distribution to the first two approaches when scaled up to the provincial level, but appears to cover a larger area, encompassing most of the entire Edmonton-Calgary corridor. As well, the nighttime light layer picks up large areas of land that are not even present in the first two RF approaches, such as the northern British Columbia / Alberta provincial border. Along the north east of the study area, there are extremely dense areas of nighttime light that are recorded, with the Fort McMurry area being captured as one contiguous mass of light (Figure 13 – C). Additionally, several industrial areas appear to have dwellings incorrectly predicted within them. In Figure 15, we see cells that have no structures present within them, other than oil well pads, and these cells were predicted to have several dwellings present within them

The differences between the multivariate and univariate approach can be seen more clearly in Figure 16, looking specifically at the Greater Edmonton Area. In the top inset, we have the RF regression approach, with the nighttime light approach beneath it, and it is apparent that the multivariate method results in a gradual decrease in predicted dwellings, as one moves further away from Edmonton. There are little clusters of predicted dwellings present around smaller cities in the region, as well as running along roadways, but we see a drop as we head into more rural areas. In the univariate approach, there is a similar large concentration seen around Edmonton, however the gradual decrease in presence is not seen, with the entire area having a homogenous coverage of recorded nighttime light (Figure 16).

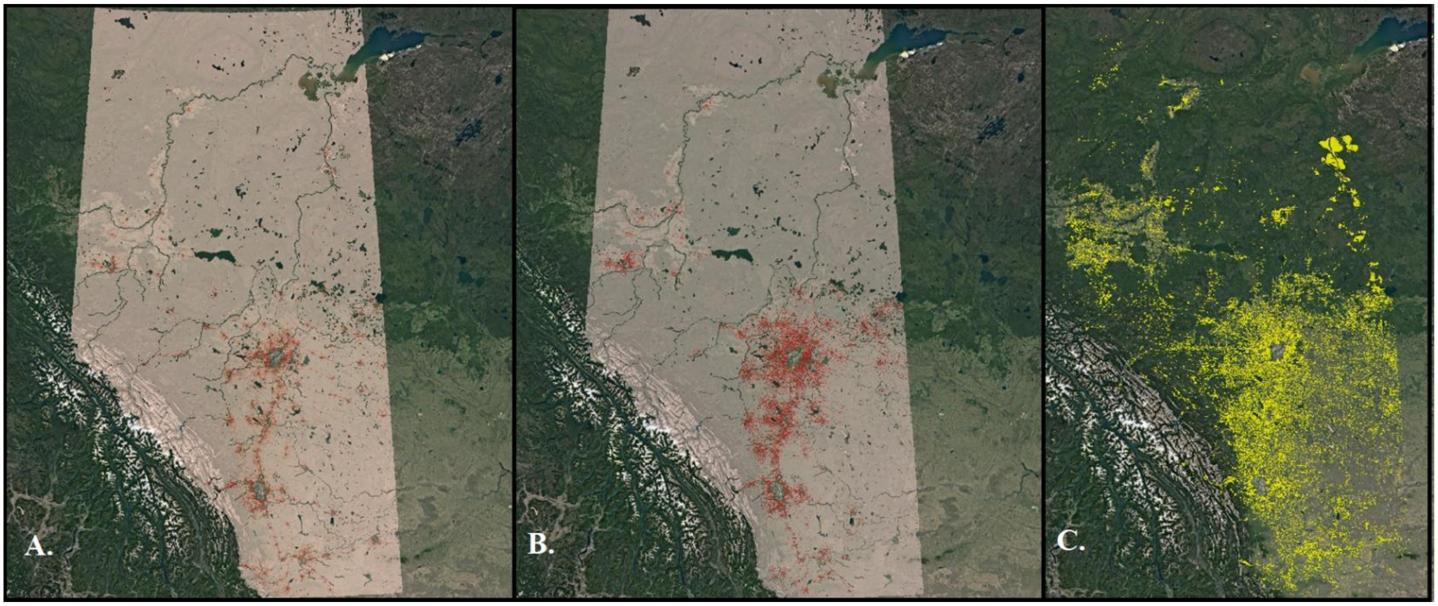


*Figure 13: Dwelling prediction approaches, scaled up to the entire study area. A: RF Regression Approach. B: RF Binary Approach. C: Nighttime Light Univariate Approach.*

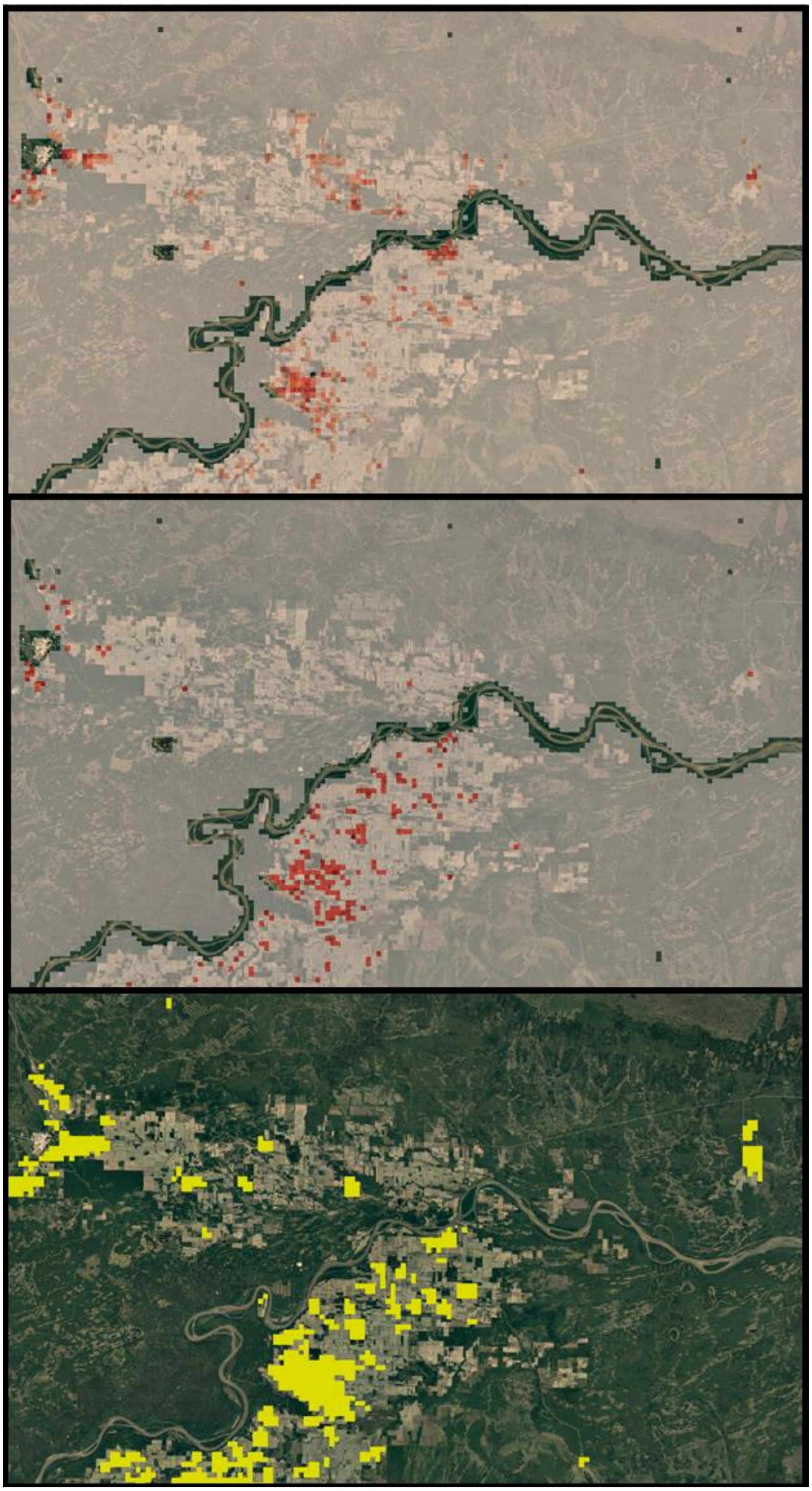


*Figure 14: (Top to Bottom) – RF Regression / RF Binary / Nighttime Light (Peace River, Alberta)*

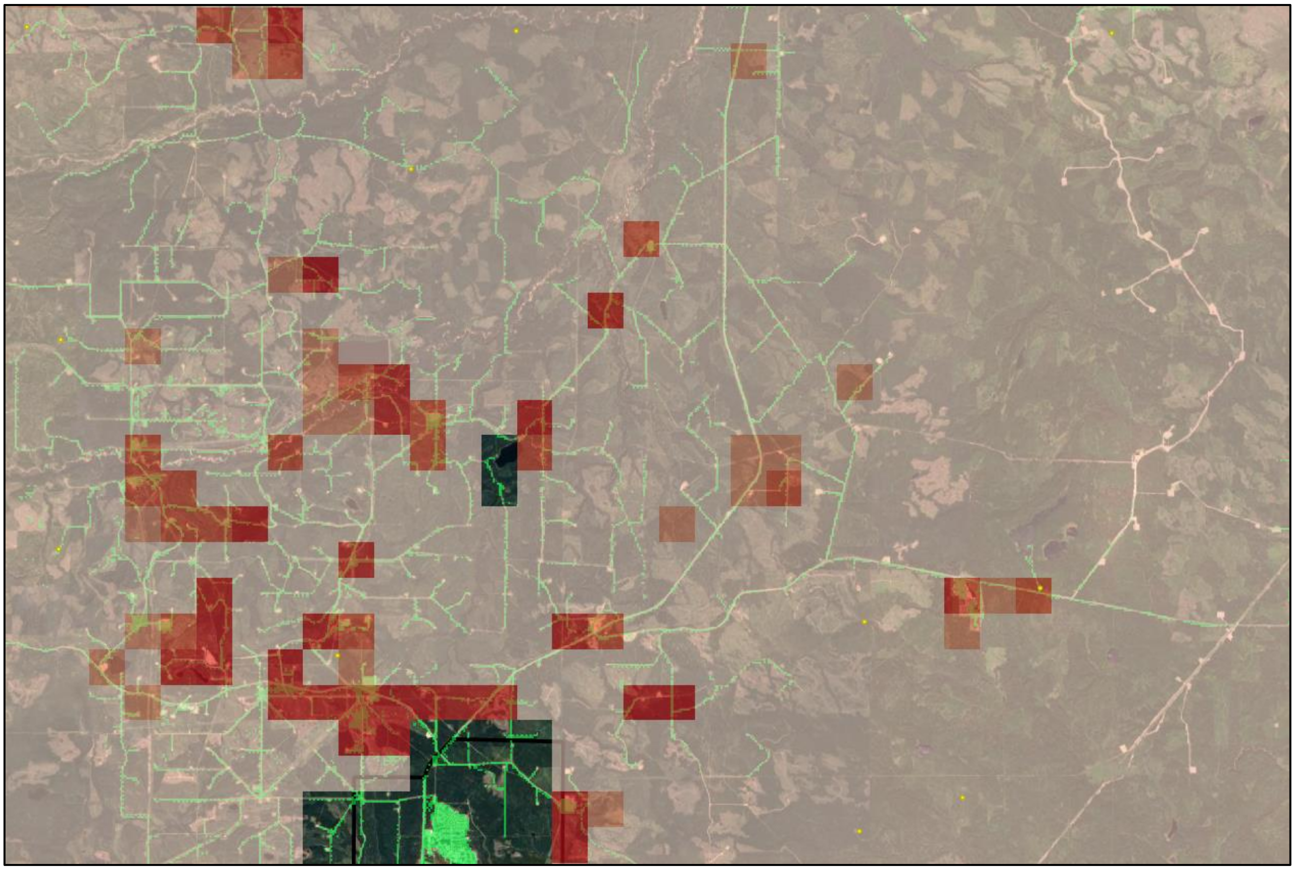


*Figure 15: Urban Land Cover (Green) Influence on the model*

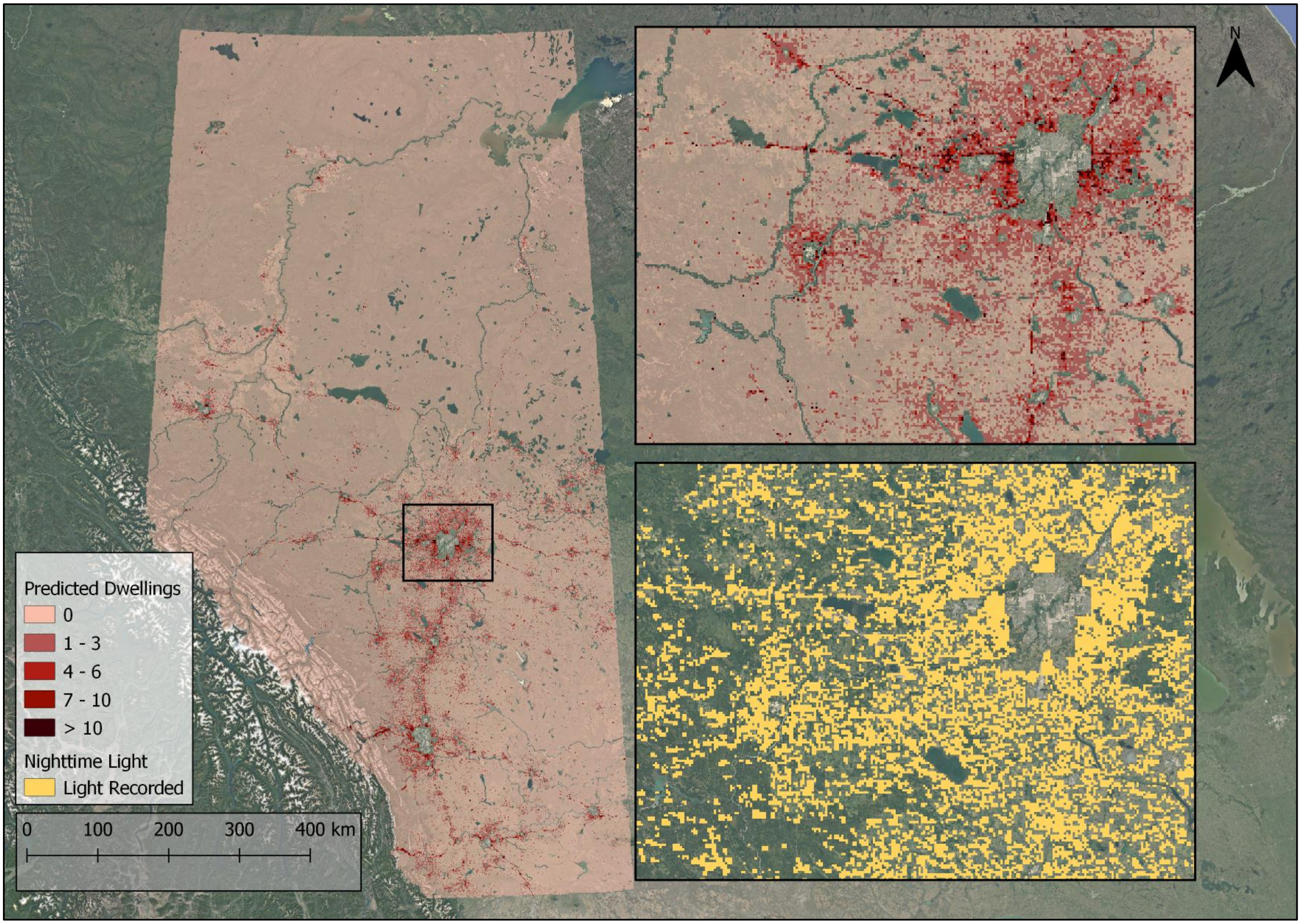


*Figure 16: Comparing the RF Regression Approach to the Univariate Nighttime Light Approach in Predicting Dwellings. Edmonton, Alberta included in inset (RF Regression approach on top, with NTL below)*

## 2.4 DISCUSSION

### 2.4.1 Summary of Findings

The aim of this study was to create and evaluate a dwelling exposure model for rural Alberta. When constructing the RDDM, we followed similar principles as LandScan, while incorporating regional spatial data related to dwelling build up. Once created, this layer can be used to create an index to disaggregate census dwelling count values. We compared this model with a simpler univariate index based on a nighttime light layer. By doing so, we wanted to understand whether this simpler approach is a feasible alternative to creating a multivariate model.

The RF regression approach produced a root mean square error value that remained relatively similar when stratified by proximity to the nearest TUC. As well, there was a slight decrease in error further away from populated areas. The RF regression model tended to under predict in areas with low dwelling counts, while still producing a distribution similar to the testing dataset. This is essential since our end goal was not to create a tool to predict final dwelling counts, but to instead create a weighted surface across the entire study area.

The RF binary layer consistently outperformed the nighttime light layer in terms of accuracy, having the higher recorded score at both the provincial and the stratified levels. In the confusion matrix, the RF binary model had higher sensitivity to identifying dwellings, compared to the nighttime light approach which had higher specificity at each distance band. The RF model underperforming in terms of specificity at each distance band, while having the highest specificity overall is likely due to it being trained at the provincial level, while the NTL data was immediately used to indicate dwelling presence. This suggests that while on the whole, the RF model outperformed the NTL data in this category, the characteristics at the distance band level favor using direct nighttime light data. With our focus being on rural Alberta and having none or few dwellings present in the majority of the photo-interpreted sites, sensitivity is arguably more important than specificity. While there is a potential for over predicting the number of dwellings exposed during a flooding event, policymakers would likely focus on minimizing the number of dwellings *not* captured by the exposure model, to ensure that planning allows for preparing sufficient resources to aid these communities in an emergency. In terms of a model's usefulness for mapping sparsely populated regions, being able to parse out the largest ratio of signal to noise is essential. The RF binary method was seen as an improvement over simply using nighttime light as an indicator of dwelling presence. With that said, using either approach to disaggregate dwelling counts would be preferred over using census polygons to understand flood exposure.

When the dwelling count model was scaled up in the more rural areas of Alberta, it was expected that the nighttime light approach would indicate the largest area covered by dwellings, as it captured all light indiscriminately and did not distinguish between dwellings and other structures. Looking at the RF regression output, we see a much lower amount of predicted dwelling counts, with the few that are present clustered into small regions near small communities.

During the photointerpretation stage, there were a few extreme outliers in terms of recorded dwellings. All these sites were in close proximity to a TUC and were often some form of dense neighbourhood or suburb. While this data was kept in the current study, it may be beneficial to remove these outliers in future modelling, as they may  overpredict dwelling counts in densely packed areas. More importantly, this points to the issue with the TUC layer itself, as there were several large suburbs that fell just outside these polygons and could have skewed the models further if they were included in the training data. Creating a buffer around the TUC layers to consider any edge effects giving outliers would solve this issue.

### 2.4.2 Limitations

Breaking the model performance down into the 3 stratified regions, the RF regression model had the highest amount of error for sites that were closest to a town, urban center, or city. For the binary performance, the RF binary model had a better sensitivity (picking up true positives) at every separated region, while the nighttime light layer performed better at specificity (showing true negatives). It is important to note that while these regions are split up in terms of their distance to a TUC, the dwelling exposure model itself was trained by data randomly collected across the entire study region. It may benefit policy makers to understand how a “rural-only” dwelling exposure model may perform in rural areas, compared to our dwelling count model. In future modelling, random cluster sampling should be explored as a way to approach these issues, while maintaining the same stratifications. Furthermore, the random sampling method used in the RDDM’s creation prohibited us from determining whether or not there was any spatial autocorrelation in both the dependent variable as well as the residuals, which would also be addressed by carrying out a form of cluster sampling.

Overall, there is an issue of generalizability when it comes to creating a regional model. For example, when it comes to creating a "rural-only" model that is only trained on sparsely populated areas, this can become costly, and may not have a significant increase in flood exposure prediction performance. Furthermore, to fully understand the value of our region-specific model in comparison to a global Human Settlement Layer, we must measure its ability to successfully disaggregate census data. Since our model was explicitly made to look at the present data gap (dwelling prediction) and not look at population counts, densities, or “all-structures counts”, a simple direct comparison is difficult. We addressed this by using the nighttime light as a benchmark, which could also be used to disaggregate dwelling counts, and then be compared the regional model.

#### 2.4.2.1 Nighttime Light Data

Nighttime light was selected to be our univariate benchmark, to compare the performances of using a variable collected at a global scale to a region-specific model in predicting dwellings. This variable has been used in several instances of measuring population, such as when Elshorbagy et al. (2017) used a combination of nighttime light and land cover classes to measure exposure, and when Anderson et al. (2010) measured the relationship between nighttime light and population density in the United States.

There are a few issues that commonly arise with using nighttime light data, such as overglow, where cells have a nighttime light value greater than what is actually present due to coarse spatial resolution, as well as overlap between pixels (Doll, 2008). In our research, we found site-specific issues with the nighttime-light-only approach, such as in dense commercial areas like Fort McMurray that had extremely high concentrations of measured nighttime light. This resulted in a dense uninterrupted mass of recorded light that is attributable to a large concentration of industrial facilities, and not a large number of dwellings. This would have a significant impact on disaggregation, with the dwelling counts in the respective administrative boundary being misallocated to these cells. Nighttime light is still an effective tool in representing economic investment and productivity, which is also important when trying to understand flood exposure. The aims of this research however is to make a tool to disaggregate dwelling counts to understand flood exposure to homes, versus making a general risk map.

To measure the performance of our RF model in comparison with the nighttime light layer, both were converted into a binary format. This resulted in a loss of detail in both models, as well as changed the analytical approach needed to evaluate the results. A potential solution to this would be to create a nighttime light index, which would separate values into categories, allowing for some of the dwelling count distribution to be possibly captured. However, this may not prove to be a large issue if the differences between these approaches are negligible once disaggregated. While simpler, the nighttime light approach requires vastly less time and resources to process and could be done much more quickly. This is essential in situations like a flooding event, where relatively coarse data can be more useful if generated quickly, in comparison to site-specific models that take valuable time to refine.

#### 2.4.2.2 Distinguishing between industrial and residential sites

While the industrial related variables such as well-pad count may aid in predicting dwelling count, it may also result in industrial areas having dwellings improperly predicted within them (Figure 15). This suggests further refinement is needed, but like the nighttime light data, may be negligible when carrying out disaggregation. The same issue is present with regards to land class cells defined as urban, which also captured roads and commercial structures, possibly impacting predictions (Figure 15). This information is still of some use though, as road density would have an impact on response time when modelling flood impact, with some remote areas potentially being more vulnerable to others due to their inaccessibility.

#### 2.4.2.3 Data Collection & Random Forest Modelling

This entire study was carried out using only publicly accessible information, which was then used to create the dwelling prediction models. The use of publicly available data means that the method is generally reproducible.  This process is dependent on online data archiving maintained by the Canadian federal and provincial governments. This precise approach is not possible where equivalent data are available.

Both the regression and binary approaches were created using a random forest framework. In terms of other machine learning approaches, both support vector machines, as well as convolutional neural networks have been used in several instances to carry out population mapping, and may be feasible alternatives (Weaver et al, 2018). The aims of this research were not to specifically focus on comparing machine learning techniques and their performance, but instead investigate whether the combination of region-specific ancillary data and collected photointerpretation data made a suitable alternative to the census. With this in mind, the random forest approach was completely acceptable.

## 2.5 CONCLUSION

This research set out to create a dwelling exposure model for rural Alberta using regional data, to then act as an index to disaggregate census dwelling values. As well, we compared our dwelling model to a single variable layer, to determine whether using a single proxy for dwelling presence would be a feasible alternative for policymakers.

The results of the random forest model to nighttime light comparison suggest that the RF model outperforms the nighttime light layer in both accuracy and sensitivity (correctly predicting dwelling presence), as well as capturing the dwelling count distribution. Both approaches use open sourced spatial data, however the random forest model required much more data, data collection, processing time and technical knowledge. In contrast, the nighttime light layer requires less data, and less technical expertise. This provides policymakers the framework to create regional dwelling models with only the use of publicly accessible data.

## REFERENCES

Anderson, S. J., Tuttle, B. T., Powell, R. L., & Sutton, P. C. (2010). Characterizing relationships between population density and nighttime imagery for Denver, Colorado: Issues of scale and representation. International Journal of Remote Sensing, 31(21), 5733–5746. https://doi.org/10.1080/01431161.2010.496798

Alberta Energy Regulator (2018). Pipelines [Vector Polyline layer]. Available from: https://www.aer.ca/providing-information/data-and-reports/maps-mapviewers-and-shapefiles.html

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal, 69(1–2), 103–117. https://doi.org/10.1007/s10708-007-9105-9

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

Center For International Earth Science Information Network (CIESIN) - Columbia University. (2016). Documentation for Gridded Population of the World, Version 4 (GPWv4). Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/h4d50jx4

Davies, J. B. (2016). Economic analysis of the costs of flooding. Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques, 41(1–2), 204–219. https://doi.org/10.1080/07011784.2015.1055804

Dobson, J., A. Bright, E., R. Coleman, P., C. Durfee, R., & A. Worley, B. (2000). LandScan: A Global Population Database for Estimating Populations at Risk. Photogrammetric Engineering and Remote Sensing, 66, 849–857.

Doll, C. N. (2008). CIESIN thematic guide to night-time light remote sensing and its applications. Center for International Earth Science Information Network of Columbia University, Palisades, NY.

Elshorbagy, A., Bharath, R., Lakhanpal, A., Ceola, S., Montanari, A., & Lindenschmidt, K.-E. (2017). Topography- and nightlight-based national flood risk assessment in

Canada. Hydrology and Earth System Sciences, 21(4), 2219–2232. https://doi.org/10.5194/hess-21-2219-2017

Esch, T., Bachofer, F., Heldens, W., Hirner, A., Marconcini, M., Palacios-Lopez, D., Roth, A., Üreyen, S., Zeidler, J., Dech, S., & Gorelick, N. (2018). Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint. Remote Sensing, 10(6), 895. https://doi.org/10.3390/rs10060895

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., & Strano, E. (2017). Breaking new ground in mapping human settlements from space – The Global Urban Footprint. ISPRS Journal of Photogrammetry and Remote Sensing, 134, 30–42. https://doi.org/10.1016/j.isprsjprs.2017.10.012

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Government of Canada (2016). 2016 Census of Canada (Provinces, Municipalities, Dissemination Areas). Available from: http://www5.statcan.gc.ca/cansim/home-accueil?lang=eng

Government of Canada (2017a). Aboriginal Lands of Canada Legislative Boundaries [Vector Polygon Layer]. Available from: https://open.canada.ca/data/en/dataset/522b07b9-78e2-4819-b736-ad9208eb1067

Government of Canada (2017b).  North American Cooperation on Energy Information [Spatial Dataset] Available from: https://open.canada.ca/data/en/dataset/57e7bc4c-680b-4640-9fa1-ded7ce186fab

Government of Canada (2019). 2015 Land Cover of Canada [Raster Layer]. Retrieved from: https://open.canada.ca/data/en/dataset/4e615eae-b90c-420b-adee-2ca35896caf6

Government of Canada (2020). Canada Energy Regulator – Provincial and Territorial Energy Profiles. Retrieved from: https://www.cer-rec.gc.ca/nrg/ntgrtd/mrkt/nrgsstmprfls/cda-eng.html

Government of Canada (2019). 2015 Land Cover of Canada [Raster Layer]. Retrieved from: https://open.canada.ca/data/en/dataset/4e615eae-b90c-420b-adee-2ca35896caf6

Hall, O., Stroh, E., & Payá, F. M. (2012). From Census to Grids: Comparing the Gridded Population of the World with Swedish Census Records. https://doi.org/10.2174/1874923201205010001

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., & Kanae, S. (2013). Global flood risk under climate change. Nature Climate Change, 3(9), 816–821. https://doi.org/10.1038/nclimate1911

Hu, W., Patel, J. H., Robert, Z.-A., Novosad, P., Asher, S., Tang, Z., Burke, M., Lobell, D., & Ermon, S. (2019). Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 353–359. https://doi.org/10.1145/3306618.3314263

Kanevski, M., Pozdnoukhov, A., & Timonin, V. (2008). Machine Learning Algorithms for GeoSpatial Data. Applications and Software Tools. 8.

Klotz, M., Kemper, T., Geiß, C., Esch, T., & Taubenböck, H. (2016). How good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from Central Europe. Remote Sensing of Environment, 178, 191–212. https://doi.org/10.1016/j.rse.2016.03.001

Lamond, J., Booth, C., Hammond, F., & Proverbs, D. (2011). Flood Hazards: Impacts and Responses for the Built Environment. CRC Press.

Letu, H., Hara, M., Yagi, H., Naoki, K., Tana, G., Nishio, F., & Shuhei, O. (2010). Estimating energy consumption from night-time DMPS/OLS imagery after correcting for saturation effects. International Journal of Remote Sensing, 31(16), 4443-4458.

Levine, J. N., Esnard, A.-M., & Sapat, A. (2007). Population Displacement and Housing Dilemmas Due to Catastrophic Disasters. Journal of Planning Literature, 22(1), 3–15. https://doi.org/10.1177/0885412207302277

Liaw & Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Liu, Q., Sutton, P., & Elvidge, C. (2011). Relationships between Nighttime Imagery and Population Density for Hong Kong. Proceedings of the Asia-Pacific Advanced Network, 31, 79. https://doi.org/10.7125/APAN.31.9

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., & Strano, E. (2019). Outlining where humans live—The World Settlement Footprint 2015. ArXiv:1910.12707 [Cs, Eess]. http://arxiv.org/abs/1910.12707

Mück, M., Klotz, M., & Taubenböck, H. (2017). Validation of the DLR Global Urban Footprint in rural areas: A case study for Burkina Faso. 2017 Joint Urban Remote Sensing Event (JURSE), 1–4. https://doi.org/10.1109/JURSE.2017.7924618

National Centers for Environmental Information (2016). VIIRS Day/Night Band Nighttime Lights [Raster Layer]. Available From: https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M. A., Ouzounis, G. K., Scavazzon, M., Soille, P., Syrris, V., & Zanchetta, L. (2013). A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(5), 2102–2131. https://doi.org/10.1109/JSTARS.2013.2271445

Province of Alberta (2010). Grassland Vegetation Inventory (GVI) Specifications. Retrieved from: http://www.albertapcf.org/rsu_docs/grassland-vegetation-inventory-specifications-5th-edition--june-29-2010-revised---november-9-2011.pdf

Province of Alberta (2016). Property - Municipal Boundaries [Vector Polygon layer]. Available from: https://open.alberta.ca/opendata/property-municipal-boundaries

Province of Alberta (2017a). Alberta ArcHydro Phase 2 [Vector Polyline layer]. Available from: https://www.alberta.ca/hydrological-data.aspx

Province of Alberta (2017b). Alberta Provincial Digital Elevation Model [Raster Layer]. Available from: https://geodiscover.alberta.ca/geoportal/rest/metadata/item/920a56abf2284816a3a6de2b6f80685c/html

Province of Alberta (2020). Municipal Government Act. Available from: https://www.qp.alberta.ca/1266.cfm?page=m26.cfm&leg_type=Acts&isbncln=9780779745739

Robert, B., & McKercher, B. W. (1992). Understanding Western Canada's Dominion Land Survey System. Extension Division, University of Saskatchewan.

Roy, D. C., & Blaschke, T. (2014). A grid-based approach for refining population data in rural areas. Journal of Geography and Regional Planning, 7(3), 47–57. https://doi.org/10.5897/JGRP2013.0409

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLOS ONE, 10(2), e0107042. https://doi.org/10.1371/journal.pone.0107042

The Canadian Disaster Database. (2018, December 21). Available from: https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. Proceedings of the National Academy of Sciences, 115(14), 3529–3537. https://doi.org/10.1073/pnas.1715305115

Weaver, J., Moore, B., Reith, A., McKee, J., & Lunga, D. (2018). A Comparison of Machine Learning Techniques to Extract Human Settlements from High Resolution Imagery. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 6412–6415. https://doi.org/10.1109/IGARSS.2018.8518528

# CHAPTER THREE: EVALUATING REGIONAL DWELLING MODEL PERFORMANCE IN COMPARISON TO GLOBAL HUMAN SETTLEMENT LAYERS

## 3.1 INTRODUCTION

### 3.1.1 Background

Decision makers must be able to understand the possible risks that their communities face. An essential part of determining these risks include understanding how many individuals could be impacted by an environmental disaster (such as a flood or industrial accident), and where these individuals are located. When such information is available, decision makers can effectively prioritize resources so that those most affected can receive timely aid. They can also use this information during the environmental assessment process, when risks to community needs to be determined prior to approving a development project.

When determining how much of a community is exposed to a hazard, accurate population data is crucial. This is often collected through a census, normally completed every five years in Canada. Census values are widely available and are aggregated into administrative boundaries at several scales. While useful, census data is not easily compatible with other spatial data, due to its values being joined to a set of polygons, rather than in a gridded format. Moreover, these data lack geographic detail, particularly in very rural areas. Indeed, in the most rural regions of Canada, precise geographic information about the precise location where rural living persons reside is unavailable. In many rural areas, postal addresses are often associated with mailboxes in town rather than a road or street reference address systems, and there is no standardized and publicly available administrative database with precise residential locations. The result is a greater uncertainty about the hazard that rural residents may experience from an environmental disaster.

Gridded population and settlement layers have been developed over the past two decades to provide a more realistic representation of where individuals may be present, particularly in regions of the world where population censuses are infrequent or unreliable, and where large populations still live in rural and remote areas. These Human Settlement Layers, or HSLs, are not restricted to only using census values, but can incorporate several inputs related to human presence. HSLs are normally one of two formats: *population-focused,* and *structure-focused*. Population-focused HSLs use census count data and then distributes it over a given region. A simple way to do this is to take the population count for a given administrative boundary, and equally distributing it within itself, using an "areal-weighting" method, such as the Gridded Population of the World (CIESIN, 2016). This process assumes a uniform distribution of people within an administrative area and does not incorporate other related spatial variables in the modelling process. A drawback to equally distributing census counts however, is that as census units increase in area, such as in rural and remote areas, there can a gradual decrease in accuracy (Doxsey-Whitfield et al., 2015). An example of another population based HSL is LandScan (Dobson, Bright, Coleman, Durfee, & Worley, 2000). Rather than provide a population count like the Gridded Population of the World, Landscan instead provides an ambient population value, referring to the number of individuals that are present within a cell on any given day. This is determined by using population count data, but instead of uniformly distributing it across an administrative boundary, values are distributed using a layer of spatial weights related to human presence (Dobson et al., 2000).

Structure-focused HSLs are designed to detect the presence of built-up areas. This approach normally does not incorporate population count data, but instead uses physically collected information, such as satellite imagery, which is then processed into a gridded layer. The Global Urban Footprint, or GUF, is an example of a 12-meter structure focused HSL, available in a binary format with cells indicating whether an area is "built up" or "not built up" (Esch et al., 2017).

### 3.1.2 Evaluating Human Settlement Layers

Regardless of the type of layer being used, evaluating HSL performance is an important step in determining how useful these tools can be to decision makers. There have been several approaches that have been used in terms of measuring an HSLs predicting ability, depending on whether it is population or structure focused. For instance, population HSLs often use census values for reference, while structural HSLs are able to carry out a more direct layer-to-layer comparison, such as through confusion matrices (Klotz et al., 2016).

When evaluating layers that visualize ambient population counts such as LandScan however, there is not a clear benchmark for evaluation. Because LandScan was made to represent a realistic depiction of population distribution over a day, and not solely focusing on the locations of individual homes, comparing its values directly to census counts would be disingenuous, and would need to consider other geographic factors such as local economic productivity, and others (Bhaduri, Bright, Coleman, & Urban, 2007). During the construction of LandScan, its validation process consisted of generating a coarser version of the HSL using data at a larger administrative unit, and then carrying out population disaggregation, with those values being compared to the more precise version of the layer (Dobson et al., 2000).

Comparisons of population HSLs have also taken place, such as by Hall, Stroh & Paya (2012), who examined the Gridded Population of the World, LandScan, The Global Rural and Urban Mapping Project (GRUMP), and an EU population model, against high resolution ground-truth census data provided by the Swedish National Registry. LandScan outperformed the other layers, although like the other HSLs, observations were made that overprediction was common in areas with already high population densities, with underprediction in more rural areas (Hall, Stroh & Paya, 2012).

When evaluating a structure-based HSLs, a ground truth layer is often created for a more direct comparison. For instance, the GUF had a regional validation carried out during its construction, using twelve 100km by 100km sites, captured through optical imagery (Esch et al., 2017). Within these sites, one thousand random samples were taken for each of the two model outputs ("built up" and "not built up"), and then were interpreted and compared to the GUF and other structural HSLs, with the GUF outperforming the other layers (Esch et al., 2017).

### 3.1.3 The Rural Dwelling Distribution Model

Both population and structure focused HSLs are useful tools in visualizing human distribution and build up and are of value to decision makers wanting to understand how a disaster event could impact their communities. With that said, there is a present gap between these two HSL types, in that neither of them specifically address dwellings. Structure-focused HSLs are constructed to detect built up areas but make no distinction between a dwelling and any other form of building. This is especially significant when needing to understand how many dwellings are exposed during something like a flooding event, versus counting every structure. By understanding the distribution of dwellings in their own communities, decision makers are able to better plan and prioritize for potential disaster events, as exposed dwellings would require resources in a much faster time frame than a storage facility, or a place of business.

To address this, we have constructed a region-specific dwelling distribution model made for rural Alberta (see chapter 2). This model is based on a random forest regression algorithm and uses several publicly available datasets associated with dwelling and population density. These datasets include proximity to the nearest Town, Urban Community or City, nighttime light intensity, land classification, infrastructure density, and proximity to resource extraction sites.

This model was trained through the photointerpretation of randomly selected sites throughout Alberta, collecting dwelling counts using earth imagery. Validation took place in two stages. First, the outputs of the random forest regression model were measured against the testing set of dwelling counts. Next, a binary version of the dwelling model outputs was measured against a univariate nighttime light layer. Once validated, the Rural Dwelling Distribution Model, or RDDM was then scaled up to the entire study area at a resolution of 830m.

### 3.1.5 Objective

The present research aims to evaluate the RDDM’s ability to accurately distribute dwellings over a given region, as well as indicate dwelling exposure, in comparison to other Human Settlement Layers. The first evaluation will measure the RDDM’s ability to accurately allocate dwellings over a given area. Then, the RDDM's ability to predict exposed dwellings will be assessed in several different event scenarios. By doing so, we aim to see how this region-specific model compares against less complex and univariate HSLs made at a global scale, and whether or not this method is of determining exposure is a feasible alternative for decision makers.

## 3.2 METHODS

### 3.2.1 Study Area

Both HSL performance evaluations take place within Alberta, Canada. The Rural Dwelling Distribution Model was constructed and trained within this province, focusing specifically on rural areas. This excluded Towns, Cities, and Urban Communities, which are defined as any areas under 1850 m$^2$ that contain a population of 1000 or more inhabitants, and 10,000 or more inhabitants, respectively (Province of Alberta, 2020). Alberta was originally selected due to its history of flooding, which includes the 2013 Calgary floods that caused approximately $6 billion dollars in damages (Davies, 2016). Additionally, there are over 400,000 kilometers of pipelines within Alberta, carrying both natural gas and crude oil (Alberta Energy Regulator, 2018). Compared to all of Canada's approximately 840,000 kilometers of pipelines, this makes up almost half, suggesting that decision makers in Alberta have an added incentive to understand the impacts of a potential pipeline spill, which also requires accurate dwelling distribution mapping.

For evaluation 1, the entire study area is used. Here, each HSLs ability to accurately allocate dwelling counts is measured at the dissemination area, or DA level, where they will be measured against the official DA dwelling counts (Figure 17). For evaluation 2, three emergency scenarios are created. Two of these are flooding events, with the final scenario being a pipeline burst. One flooding scenario site surrounds the Red Deer river, and is southwest of the city of Red Deer, with the remaining two scenario taking place at the same site, surrounding the Elbow River, west of Calgary (Figure 18). The boundaries for both sites were determined by selecting dissemination areas that follow both the Red Deer and Elbow river, respectively.

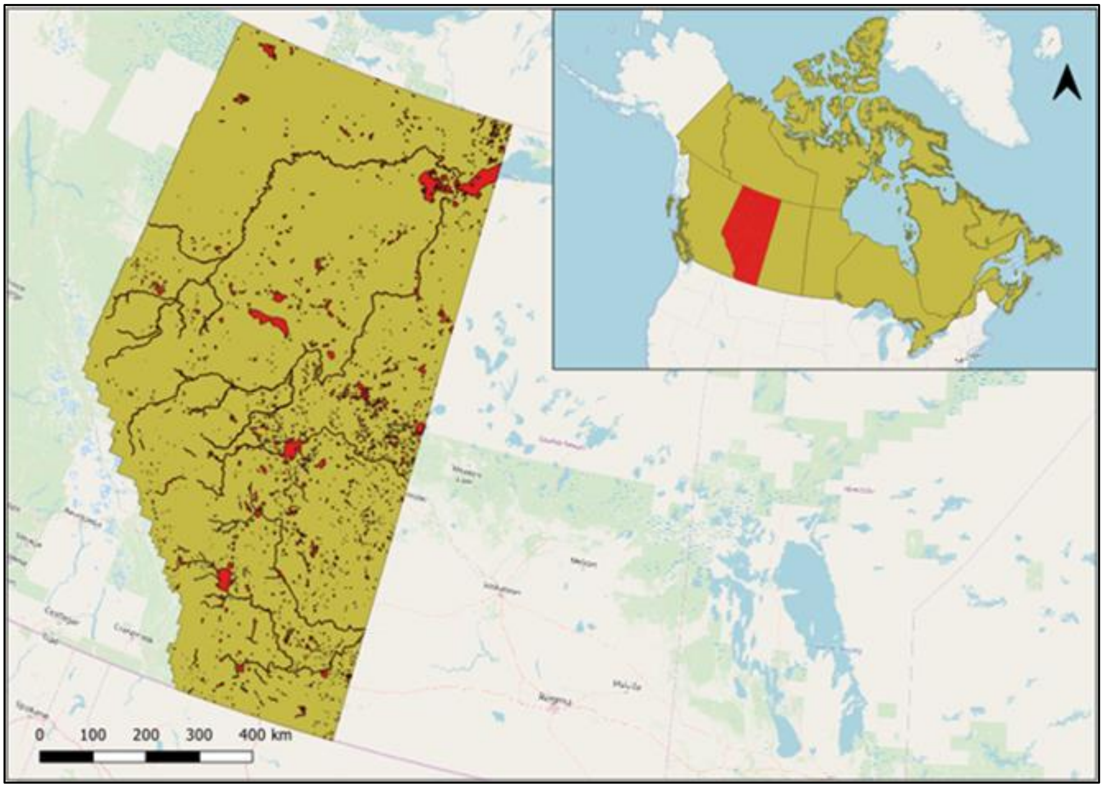


*Figure 17 : The extent of the study area (khaki) with areas in red indicating removed sections*

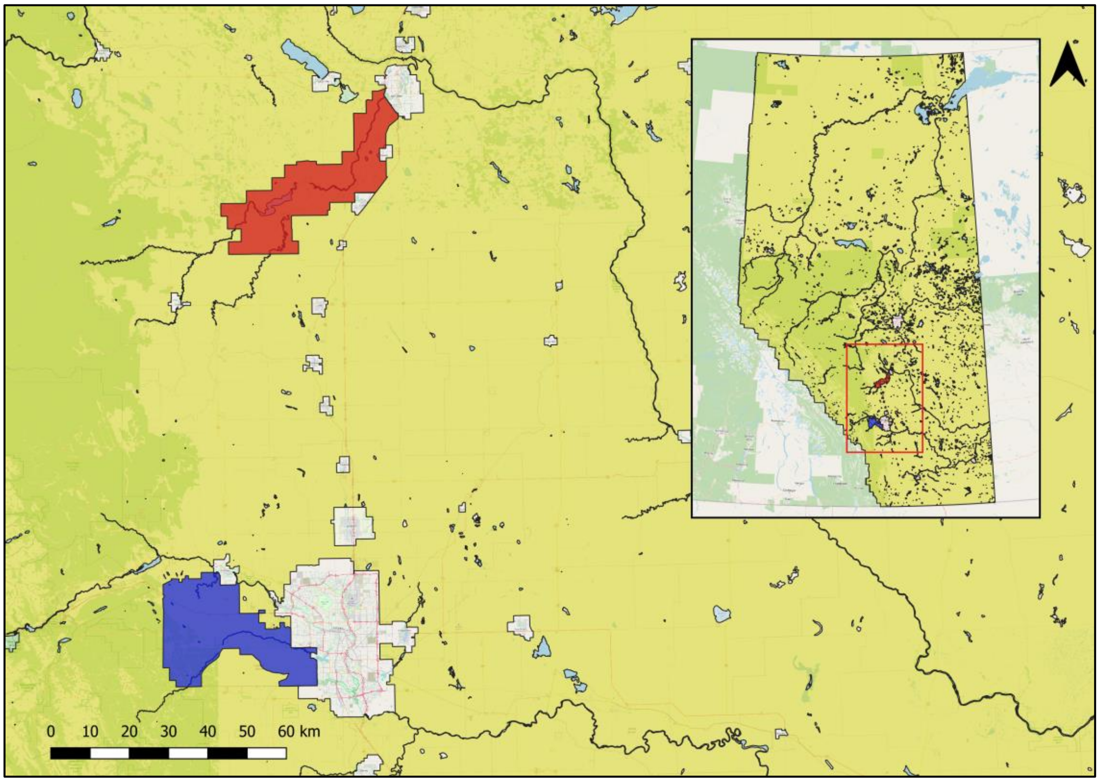


*Figure 18:Evaluation 2 Scenario Sites. Red = Red Deer Site, Blue = Calgary Site*

### 3.2.2 Data

*Census Data and Administrative Boundaries*

For evaluations 1 and 2, Canadian dwelling counts are used as a reference dataset. Defined as the total number of living quarters present within a set of administrative boundaries, this information was collected from the 2016 Canadian census (Government of Canada, 2016). Dwelling counts are collected at the Dissemination Area and Census Division, or CD, levels for Alberta, which are the smallest and third smallest publicly available census units, respectively. These census units are publicly available as vector polygons and were made accessible by Statistics Canada. Due to our in-house model only focusing on rural areas, Dissemination Areas that fell within any Town, City, or Urban Community boundary were removed, leaving 1240 Dissemination Areas.  (Figure 19).

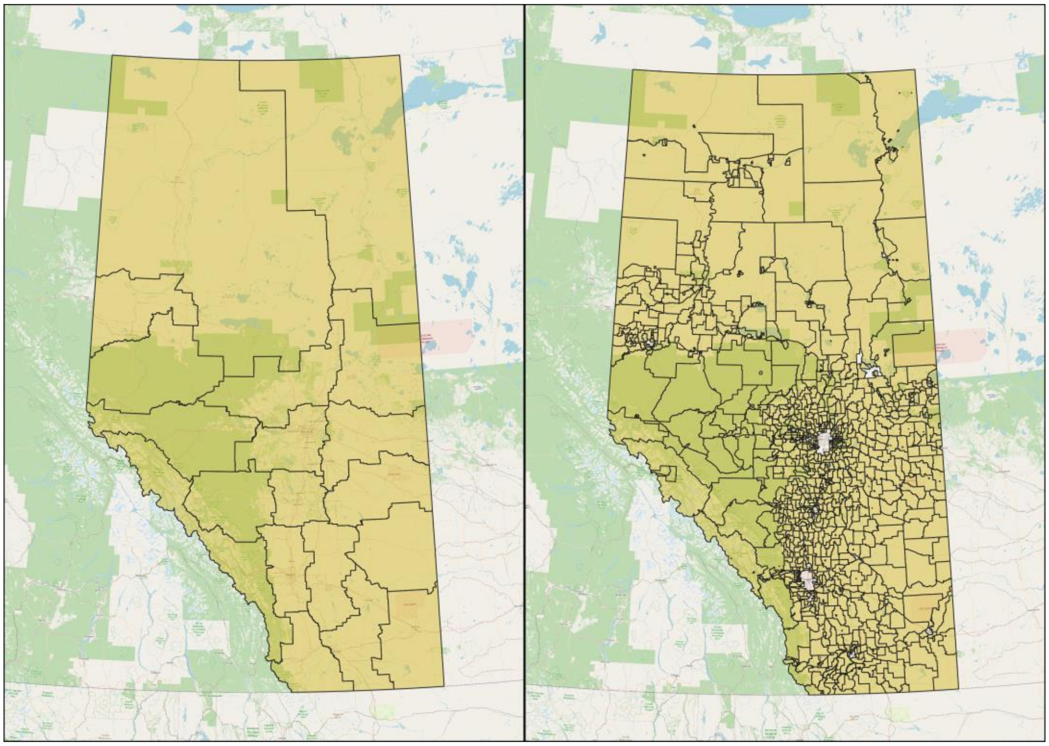


*Figure 19:Left = Alberta Census Divisions (19 ). Right = Alberta Dissemination Areas, with TUC DA's removed (1240).*

*Flooding and Pipeline Spill Scenario Data*

A vector layer of waterbodies within Alberta was retrieved from a publicly available spatial database (Province of Alberta, 2017)  From this layer, 2 sets of polygons were created for the Red Deer River and Elbow River sites. Next, a 500m buffer was generated for each set of waterbody polygons, with the total area of the buffers representing flood extent. A similar approach was used to create the pipeline spill scenario. Using an oil and gas pipeline layer made available by the Alberta Energy Regulator, an Elbow River specific clip of the layer was made, with a 500m buffer generated for all pipelines within the scenario site (Alberta Energy Regulator, 2018).

### 3.2.3 Human Settlement Layers

There are six HSLs being compared to one another in both evaluations 1 and 2. These include: the RDDM, an areal weighted layer constructed using census divisions, the Global Urban Footprint at both 12m and 830m resolutions, and a nighttime light layer, in both binary and ranked formats.

*Rural Dwelling Distribution Model (RDDM)*

The RDDM was created using 12,500 randomly selected points in the study area, using a random forest regression algorithm. Once tested, the RDDM was scaled up to the entire study area and outputted at a resolution of 830m. The current layer's cells do not represent a  final set of dwelling counts, but is instead a continuous index, which will then be used to distribute the dwelling counts provided by the census, using the same principles of LandScan (Dobson et al., 2000).

*Global Urban Footprint*

The Global Urban Footprint, or GUF, is a structure focused HSL, created by the German Aerospace Center (DLR). This layer is a binary surface, indicating where an area is "built up" or "not built up" with regards to human-made structures. The GUF was constructed using radar imagery via DLR-Satellites, which was then processed through unsupervised classification, as well as through several mask layers including water bodies and infrastructure networks (Esch et al., 2017). This layer is available at a 12-meter resolution, and the original format was retained for evaluation purposes, along with a second GUF layer resampled to 830 meters.

*Nighttime Light*

Two dwelling estimate layers were created using solely nighttime light data, produced by the Earth Observations Group at the National Oceanic and Atmospheric Administration (NOAA). This dataset contained the temporal average for 2016, at a resolution of 15 arc-seconds, which is approximately a 272-meter resolution for Alberta (National Centers for Environmental Information, 2016).

This layer was used as an input in the construction of the RDDM, and in this study, both versions of the nighttime light layer are resampled to an 830-meter resolution (same as the RDDM). The first nighttime light layer is in a binary format, where a value of "1" is assigned if any light is present within a cell, and "0" otherwise. The second nighttime light layer uses a ranking system to represent the average intensity of nighttime light present within each cell. This is similar to the approach Elshorbagy et al (2017) used when creating a composite flood exposure across all of Canada, which incorporated ranked nighttime light values and land classification. Table 7 presents each nighttime light category, and its assigned value.

*Table 7: Ranked Nighttime Light Index Categories*

| Nighttime Light Value | Assigned Value | Level |
|---|---|---|
| 0 | 0 | No Light Recorded |
| 0 - 2 | 1 | Some Light Recorded |
| 2 - 5 | 2 | Moderate Light Recorded |
| 5 - 15 | 3 | High Amount of light recorded |
| > 15 | 4 | Very High Amount of light recorded |

### *Areal Weighted Census Dwelling Layer*

Lastly, a dwelling count layer was constructed using Canadian census values, and census administrative boundaries. Using the same principles as the Gridded Population of the World, the total number of dwellings present in each census division within the study area was equally distributed within itself and was then outputted at a resolution of 830 meters.

### 3.2.4 Evaluating the Human Settlement Layers

*Evaluation 1 - Allocating Dwelling Count Values*

All 6 HSLs were evaluated for their ability to disaggregate census dwelling counts at the dissemination area level. Our aim is to determine how of the each 4 approaches (the RDDM, the areal weighted layer, both of the Global Urban Footprint layers, and both of the nighttime light layers) compare in terms of allocating dwelling counts to small regions (DAs) within Alberta, as well as any spatial allocation trends present across the layers.

Beginning with the Rural Dwelling Distribution Model, each HSL was clipped and separated by the census division boundaries within the study area. Each separate census division is then normalized by taking each of its cells values and then dividing it by the sum of the census division clip itself, so that each CD equals a value of 1 when summed. Once normalized, each CD raster is then multiplied by the total number of dwellings present in each census division, which are then distributed within the census division. Since Towns, Urban Communities, and Cities are outside the study area, all dwellings that fell within a TUC were removed from the disaggregation process. This is entire process is visualized in figure 20.

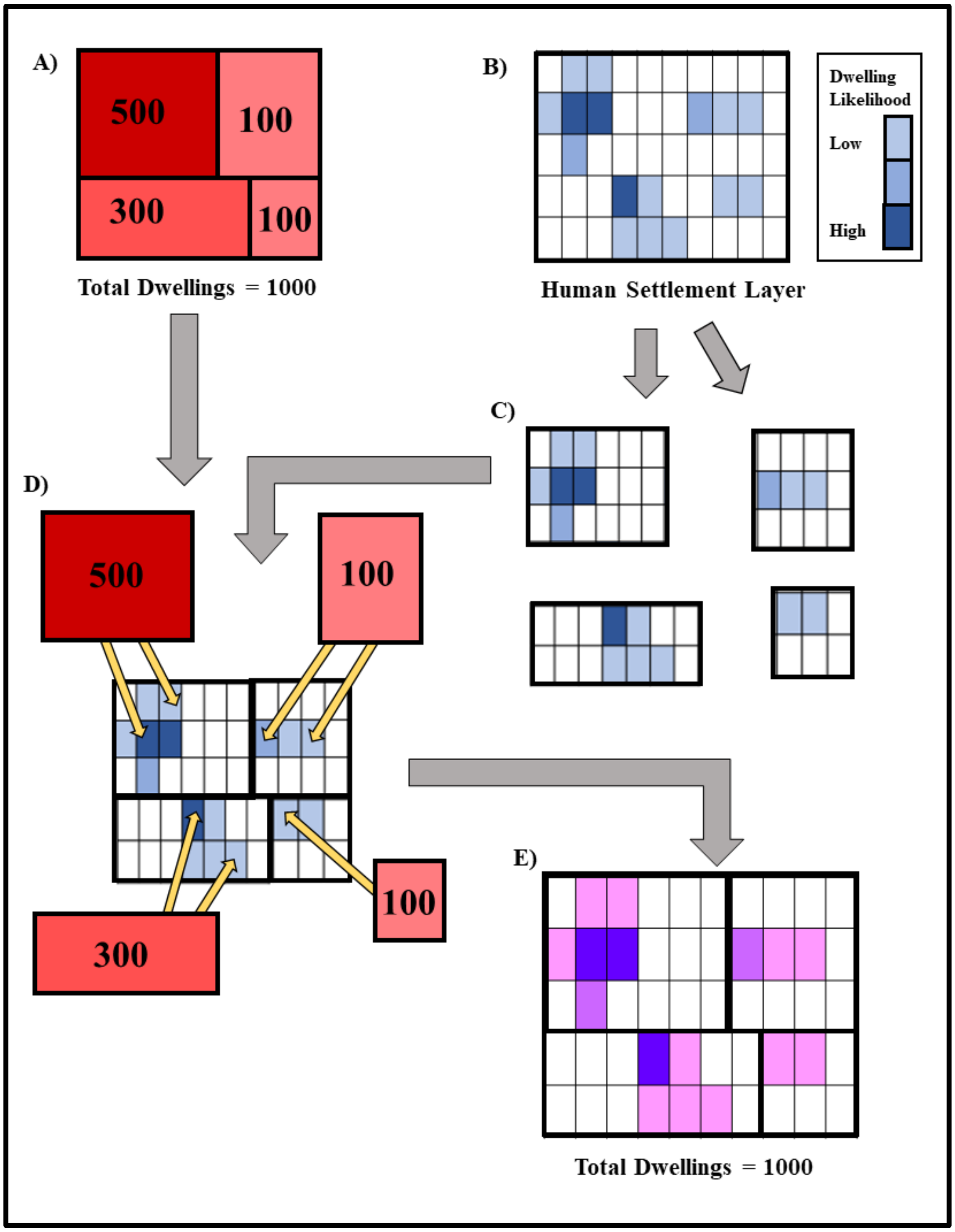


*Figure 20: The Dwelling Count Disaggregation Process – A)The Census Division Boundaries B) A Human Settlement Layer, with an index representing the likelihood of existing dwellings C) The HSL being split by CD boundaries, and then normalized D) The actual disaggregation where the split HSL is multiplied by the total number of dwellings present E) The re-merged HSL, now representing real world dwelling count values*

Once each census division is normalized, the sum of each CD raster will not exceed the actual number of dwellings present. The census divisions are then re-merged into one layer, and the total count of dwellings per *dissemination area* is recorded. Since the dissemination areas are

2 census unit sizes smaller than the census division, we are able to compare how each layer performs in allocating dwelling counts. This was completed by dividing the allocated DA dwelling count sum by the actual DA dwelling count sum. Quotients that fall between 0.75 to 1.25 were considered as acceptable, with values less than 0.25 and more than 1.75 deemed as greatly under and over allocating, respectively.

The same process was carried out with the Global Urban Footprint layers, as well as the nighttime light layers. A key difference, however, is that due to both GUF layers and one nighttime light layer being binary, there is not a continuous distribution of dwellings within the DA. Instead, each cell that has a "1" assigned to it was allocated the same number of dwellings as all other cells with an assigned "1" in the census division. Once constructed, the percentages of how many DA's fell into which allocation performance category were created for each HSL.

*Evaluation 2 - Dwelling Exposure Scenarios*

Once the disaggregation evaluations are complete, the second evaluation places each HSL in 3 different disaster scenarios to understand how each layer behaves with regards to determining dwelling exposure. As Alberta has a history of flooding, 2 scenarios involve a hypothetical flooding event, with the third consisting of a hypothetical pipeline spill.

Dwelling exposure was measured by taking the 6 (now disaggregated) HSLs and clipping them using the dissemination area boundaries shown in Figure 18. Once clipped, the total number of allocated dwellings that fall within each of the three buffers are calculated for each layer, which is then divided against the total number of dwellings within the scenario boundaries. The resulting value would provide a percentage of how many dwellings in the selected DA's are deemed "exposed". Along with the 5 HSLs being compared, the areal weighted dwelling count layer is also included in the dwelling exposure exercise. This layer is constructed using original dwelling count values, with each census division having the total number of dwellings within it equally distributed across each cell, similar to the Gridded Population of the World (CIESIN, 2016). Rather than use a more complex approach to simulating a flooding event or a pipeline spill, this method was chosen as it is quite generalizable, and can act as a valuable representation of each layer's tendencies towards liberal or conservative dwelling exposure estimates.

## 3.3 RESULTS

### 3.3.1 Evaluation 1 – Dwelling Count Allocation Performance

Figures 21-24 present each layer's dwelling allocation performance after disaggregating at the census division level. In table 8 we see each HSL's allocation performance by percent, indicating the amount of DA's that fell into each allocation category. Out of the 6 approaches, the 830m GUF layer had the worst performance in terms of allocating dwellings within an acceptable range. Conversely, the RDDM was the best performing, with just over 27% of its DA's having an acceptable range of allocated dwellings, with the 12m GUF layer having the second-best performance, with approximately 23% of its DA dwelling counts being properly allocated.

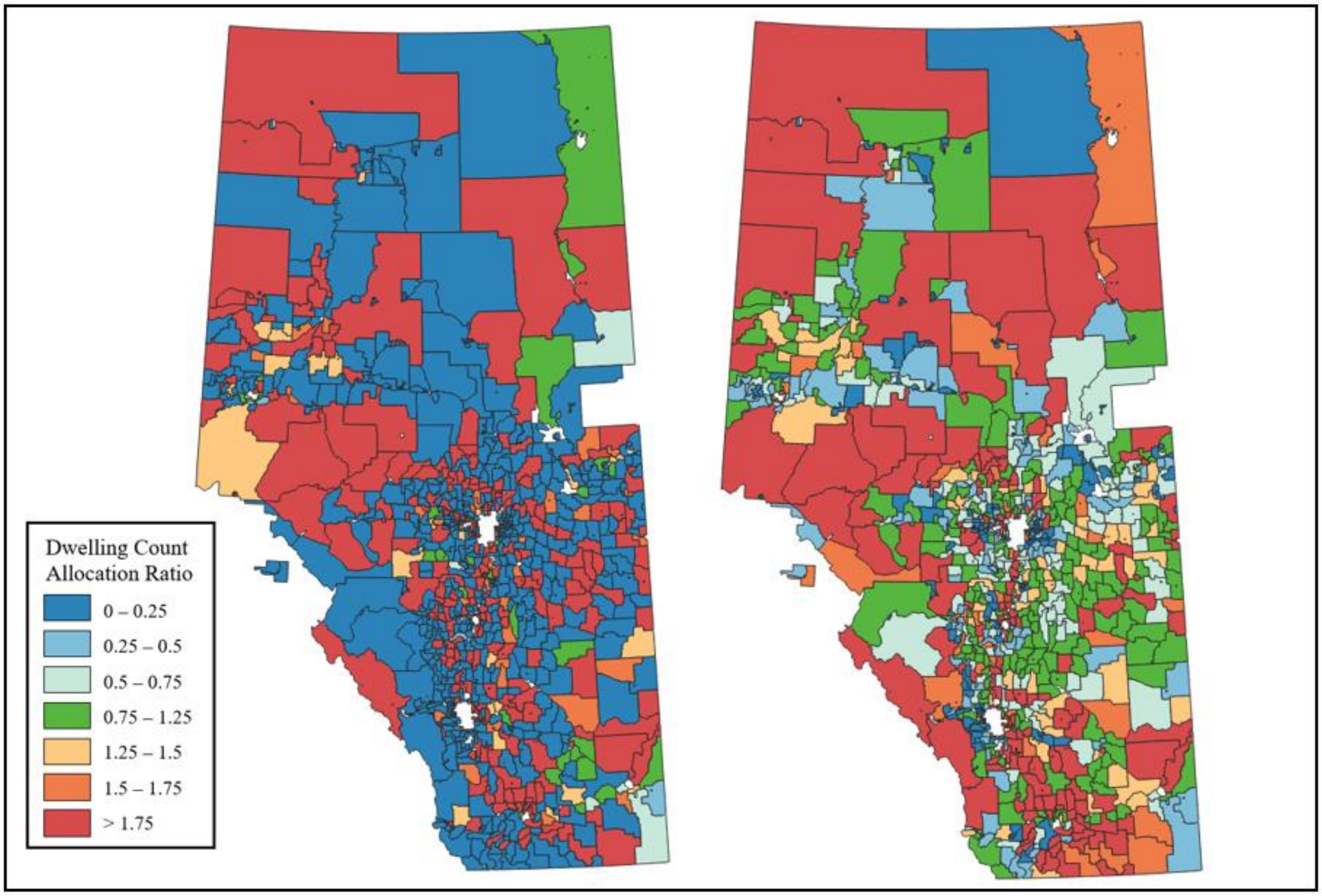


*Figure 21: Global Urban Footprint Dwelling Allocation. Left = 830m Resolution. Right = 12m Resolution*

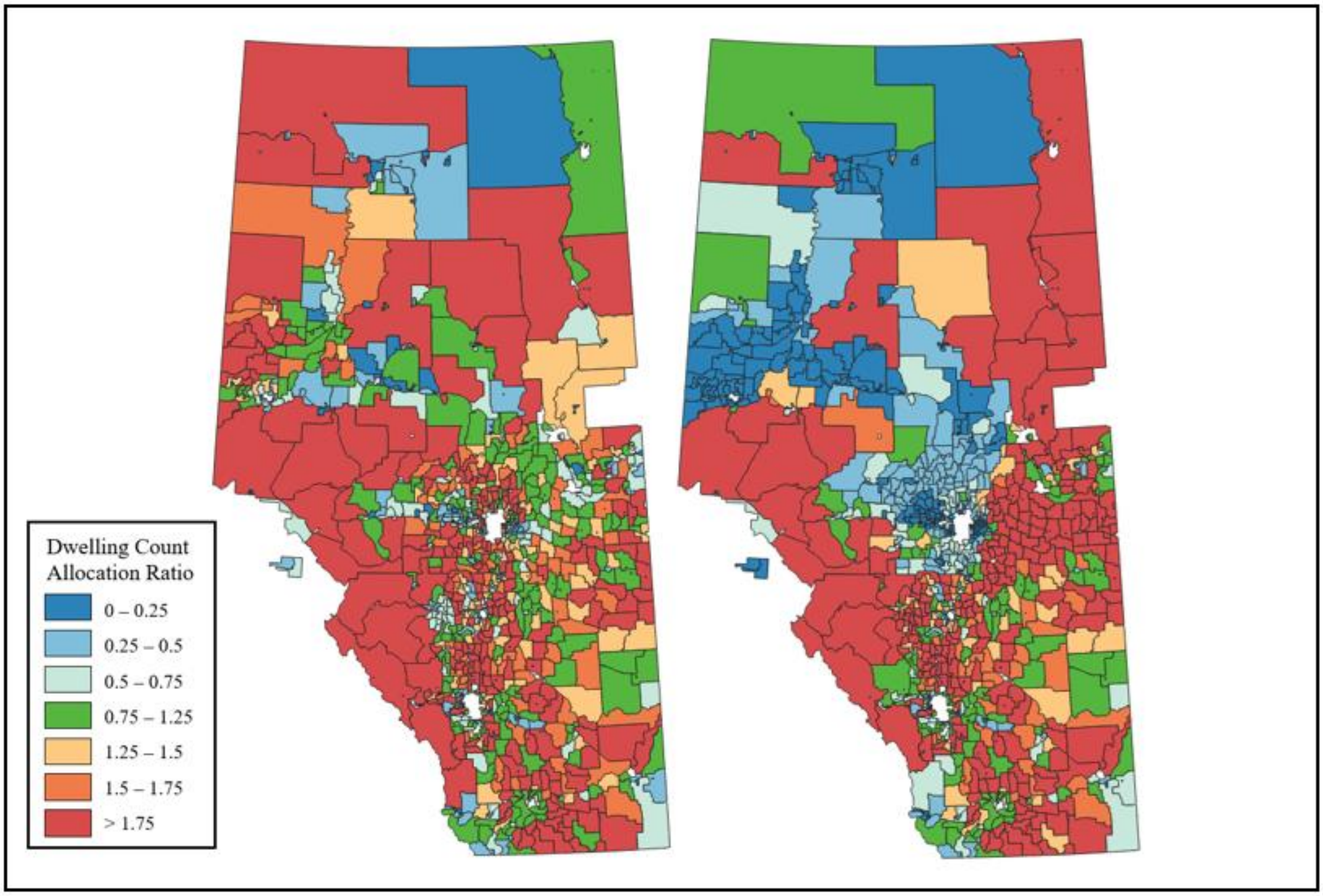


*Figure 22: Nighttime Light Dwelling Allocation. Left = Indexed Layer. Right = Binary Layer*

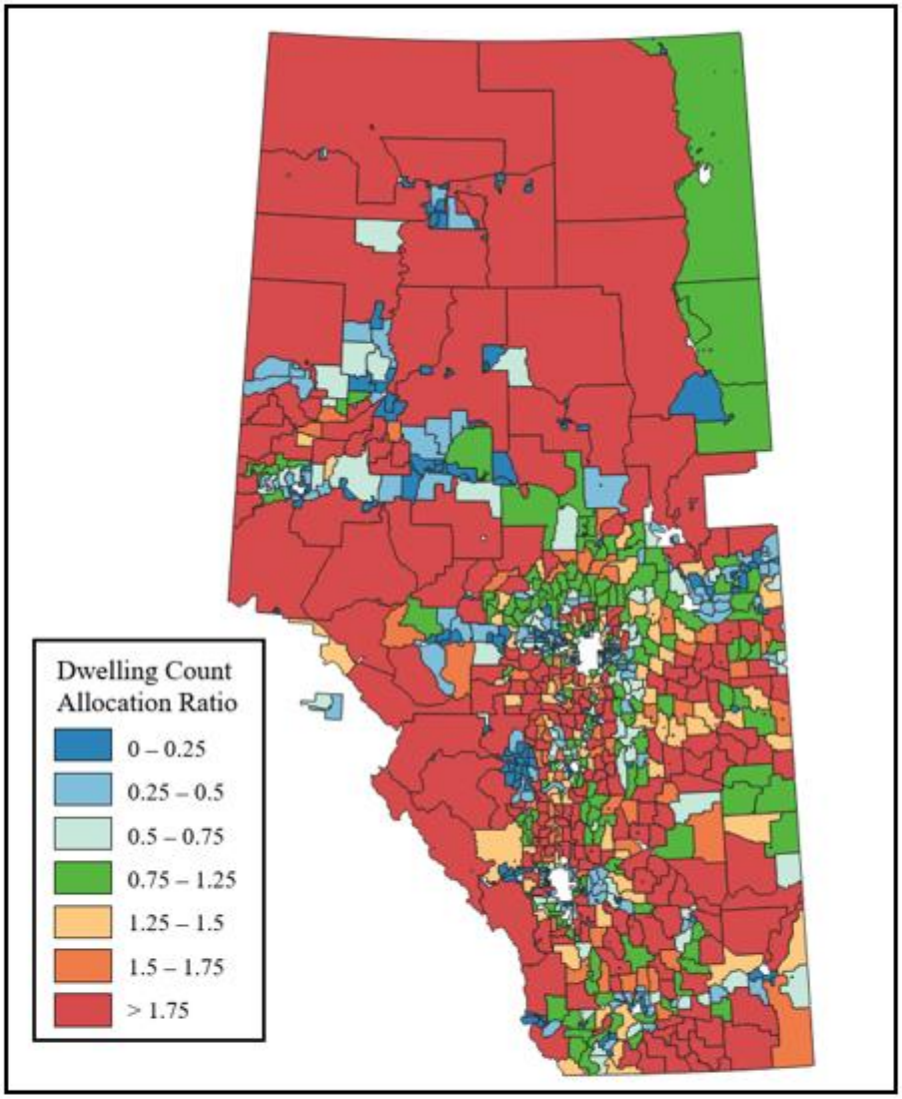


*Figure 23: Areal Weighted Layer Dwelling Allocation*

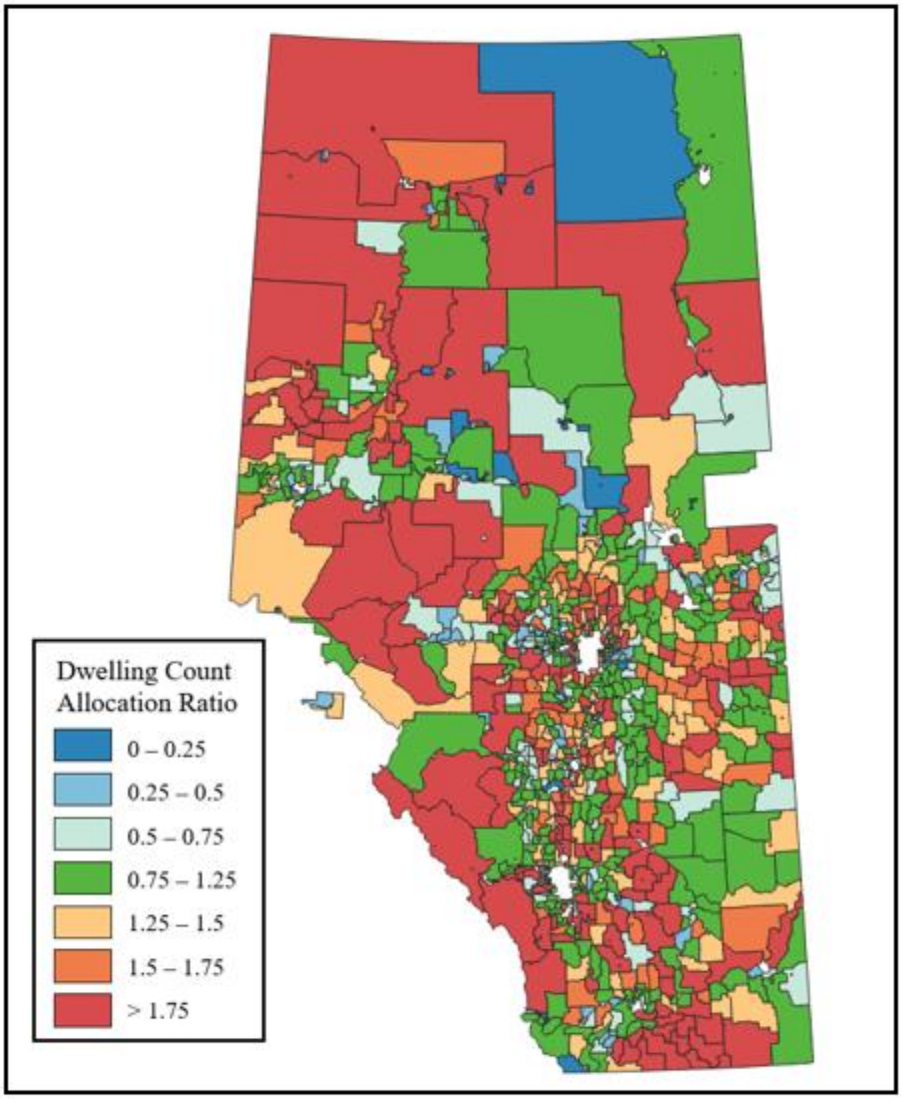


*Figure 24: Rural Dwelling Distribution Model Dwelling Allocation*

*Table 8: Dwelling Count Allocation Results (Total % of each Layer)*

| **Dwelling Count Allocation Ratio** | **NTL Index** | **NTL Binary** | **Areal Weighted** | **GUF 830m** | **GUF 12m** | **Rural Dwelling Distribution Model** |
|---|---|---|---|---|---|---|
| **0-0.25** | 26.03 | 40.34 | 35.94 | **65.74** | 21.16 | 16.29 |
| **0.25-0.5** | 11.08 | 12.35 | 12.43 | 0.84 | 16.20 | 11.42 |
| **0.5-0.75** | 11.00 | 8.07 | 8.65 | 1.26 | 15.62 | 13.77 |
| **0.75-1.25** | 17.55 | 10.92 | 13.52 | 3.44 | **23.43** | **27.04** |
| **1.25-1.5** | 6.97 | 4.37 | 6.05 | 2.85 | 6.21 | 9.57 |
| **1.5-1.75** | 8.31 | 4.20 | 5.21 | 3.27 | 3.61 | 7.14 |
| **> 1.75** | 19.06 | 19.75 | 18.22 | **22.59** | 13.77 | 14.78 |

Both nighttime light layers, the area weighted layer, and the 830m GUF layer all tended to under allocate dwelling counts across the study area, with the 830m GUF and binary nighttime light layers having the largest amounts of under allocated DA's. This was not the case for the 12m GUF layer and the RDDM, which had the lowest amount of under allocated dwellings. Spatially, the 830m GUF layer has its under allocated DA's spread out across Alberta, while the binary nighttime light layer has the majority of its under allocated DA's concentrated northwest of Edmonton, Alberta's capital (Figure 21-24). In terms of over allocating dwelling counts, the binary nighttime light layer has large concentrations located on the western and eastern borders of the province, along with the areal weighted layer, having most of its larger and more remote DA’s over allocated.

Table 9 presents the mean absolute error (MAE) for each HSL with regards to allocated dwelling count versus the official census value. Here we see that the RDDM possessed the lowest MAE, followed by the 12m GUF layer, and the Indexed NTL layer. Similar to Table 8, the 830m GUF layer and the Binary NTL had the highest MAE values, with 258.13 and 202.09, respectively.

*Table 9: Human Settlement Layer Mean Absolute Error*

| HSL Mean Absolute Error | NTL Binary | NTL Indexed | GUF 830m | GUF 12m | Areal Weighted Layer | Rural Dwelling Distribution Model |
|---|---|---|---|---|---|---|
| | **202.09** | **149.63** | **258.13** | **140.70** | **190.86** | **115.87** |

There was a consistently over allocated dissemination area seen in all 6 layers, at the southwestern border of Alberta and British Columbia. Upon further inspection, high dwelling counts were observed around the town of Jasper, and using earth imagery data, cells with high amounts of allocated dwellings overlapped with a recreational vehicle campground. Figure 25.1-25.6 presents these findings, using the indexed nighttime light layer as a reference.

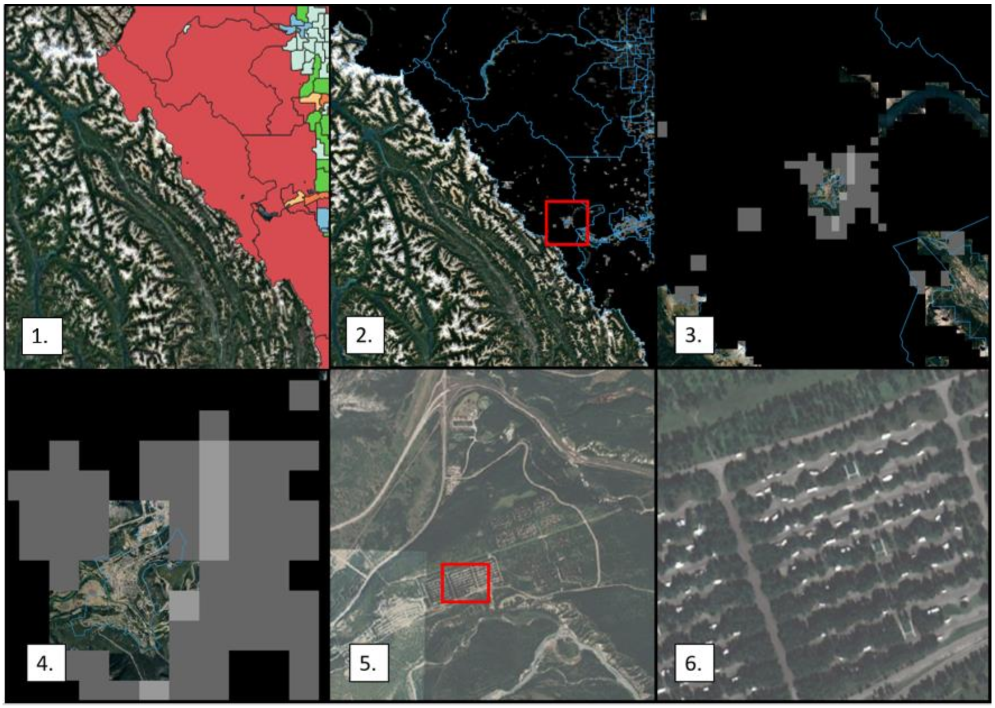


*Figure 25:Dwelling Over Allocation (Near Jasper).*

### 3.3.2 Evaluation 2 – Scenario Sites

Figure 26 presents the 3 scenarios, with the Calgary flood and pipeline spill scenarios on the left, and the Red Deer flood scenario on the right most square. All 5 disaggregated layers, along with the areal weighted dwelling count layer were clipped to these boundaries, and the total number of dwellings that fell within the buffered polygons were divided by the total number of allocated dwellings within the scenario boundaries. The resulting values provide an approximation of how many dwellings are "exposed" in each event.

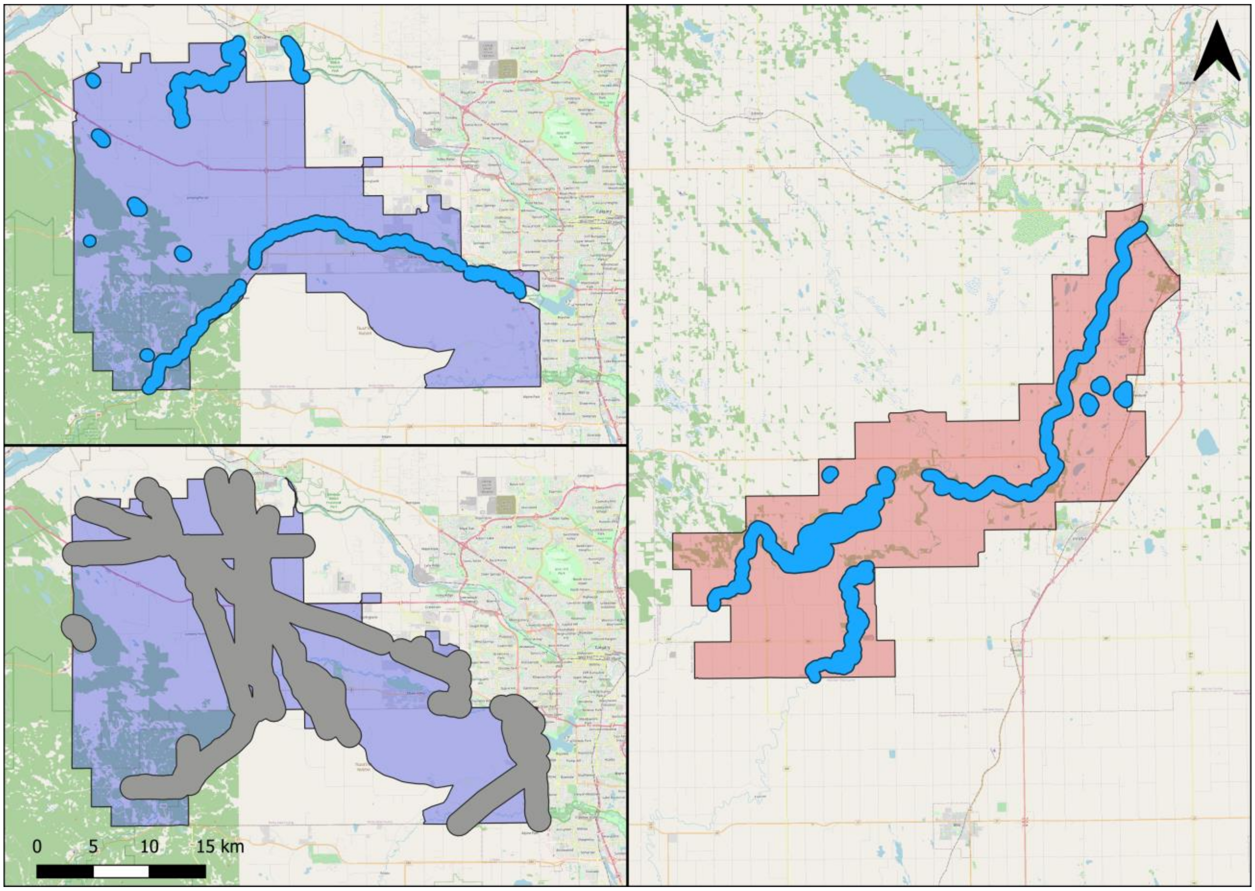


*Figure 26: Event Scenario Sites. Upper Left = Calgary Flooding Event. Lower Left = Calgary Pipeline Spill. Right= Red Deer Flooding Event.*

Table 10 presents the scenario results for all 6 layers. Here we can see the relative amount of dwellings each layer predicts would be exposed In all three scenarios, the RDDM provided the lowest estimates in terms of exposed dwellings, with the 830m GUF layer providing the highest amount of exposed dwellings in each scenario.

*Table 10: Dwelling Exposure by HSL (% exposed dwellings with respect to scenario site)*

| Location / Scenario | NTL Binary | NTL Index | GUF 830m | GUF 12m | Rural Dwelling Distribution Model | Areal Weighted Dwelling Counts |
|---|---|---|---|---|---|---|
| **Calgary Flooding Event** | **5.93** | **6.19** | **66.67** | **28.7** | **5.2** | **10.34** |
| **Red Deer Flooding Event** | **4.23** | **4.32** | **33.33** | **25.47** | **3.4** | **15.05** |
| **Calgary Pipeline Spill Event** | **31.03** | **32.45** | **66.67** | **41.6** | **29.94** | **26.98** |

## 3.4 DISCUSSION

The goal of this research was to both evaluate the Rural Dwelling Distribution Model's ability to accurate allocate dwelling counts over a given region, as well as understand its behaviour when predicting exposed dwellings, given various disaster scenarios. This evaluation included several other Human Settlement Layers, to provide alternative approaches to decision makers wanting to understand the risks facing their communities. By doing so, we wanted to understand how the RDDM, a region-specific layer, compared against less complex and univariate HSLs made at a global scale.

### 3.4.1 Dwelling Count Allocation

When compared against the census dwelling counts, the RDDM and the 12m GUF layer had the two highest amounts of dissemination areas that fell within the acceptable allocation range, as well as the lowest mean absolute error (Table 8-9). Conversely, the 830m GUF layer performed the poorest out of the 6 layers, with both the lowest number of DAs within the acceptably allocated range, as well as having the highest amount of error. The areal weighted layer was outperformed by the indexed nighttime light layer, the 12m GUF layer, and the RDDM. As expected, the areal weighted layer tended to over allocate dwellings in large DAs, due to its equal distribution of dwellings across entire census divisions. By doing so, population centers within each census division were greatly under allocated, with the remaining dwellings being allocated to extremely rural and remote regions of the province.

With regards to the two nighttime light layers, the indexed layer outperformed its binary counterpart in terms of allocating dwelling counts, with the binary NTL layer tending to under allocate dwellings across the DAs. This trend continues with the Binary NTL layer having a higher amount of error when compared to its ranked counterpart, suggesting that if decision makers are forced to use a univariate layer to allocate dwelling counts in a region, using a continuous layer would be the better performing tool. This is presumably due to the fact that unlike the RDDM and the indexed NTL layer, dwelling count values are distributed equally to every cell indicating a structure in a binary layer, leading to an inability to accurately allocate dwelling distributions. Notably, this drop in accuracy was not present in every binary layer, with the 12m GUF layer having the second-best performance of the five. This is most likely due to the relatively high resolution of this layer in comparison to the other four.

Lastly, all six layers had consistently over allocated dwellings to a dissemination area in the Jasper region of Alberta, near its southwest border with British Columbia (Figure 25). Upon further examination, it was seen that this DA had several recreational vehicle campgrounds within it, while also having a relatively low number of census recorded dwellings, leading to this outcome. While technically inaccurate with respect to the census values, this scenario can still be of use to decision makers when planning for disaster events, as these sites were still populated, however only on a seasonal basis.

### 3.4.2 Scenario Site Evaluation

In all three scenarios, the RDDM consistently predicted the lowest amount of exposed dwellings. Both NTL layers predicted slightly higher amounts of exposed dwellings, followed by the Areal Weighted Census Division layer. The two GUF layers predicted the highest amount of exposed dwellings, with the 830m layer have the higher predictions of the two.

It must be noted that the while the RDDM and the 12m GUF layers performed the best in terms of allocating dwelling counts in the first evaluation, they predicted the lowest and second highest amounts of exposed dwellings, respectively. This suggests that as we move from the regional to the site-specific level, there is not a clear relationship between a layer’s dwelling count allocation performance, and its tendency to predict a relatively high or low amount of exposed dwellings. This is further suggested when looking at the nighttime light layers, where there is only a slight change in the predicted number of exposed dwellings when going from the binary NTL layer to the Indexed layer, unlike the pronounced change seen during the dwelling count allocation.

Even without a clear relationship between each layer’s dwelling count allocation performance and its behavior when predicting exposed dwellings, these results can still be of value to decision makers when deciding between layers to help determine risk. Furthermore, with the Areal Weighted layer predicting an intermediate amount of exposed dwellings in comparison to both the GUF, NTL, and RDDM layers, decision makers can use it as an alternative to modifying or creating more complex layers when determining and communicating risk.

### 3.4.3 Limitations

The three primary layers being used during this research, the RDDM, the GUF, and the Nighttime Light Layer, were all constructed in significantly different ways. The RDDM is a multivariate dwelling exposure model created using a random forest regression algorithm, the GUF layer uses high precision satellite imagery and providing a binary "all-surfaces" output, and the NTL layer specifically uses captured light. With that in mind, it would be disingenuous to immediately compare the performances of the three layers to one another, and not consider these differences.

This is especially the case when comparing the binary layers to the non-binary HSLs, since the layers equally distribute their respective dwelling counts to every cell indicating a present structure or captured light. This does not occur with the non-binary HSLs, and as such they are able to more accurately allocate dwellings within a given region. The 12m GUF layer was the one exception to this. This layer was specifically kept at its original resolution, while its 830m counterpart had the worst performance in the first set of evaluations. In future analysis, it would be beneficial to examine additional non-binary and binary HSLs and compare both sets separately, to better understand their behaviour when allocating dwellings and predicting exposure.

There are several tradeoffs to keep in mind when selecting any of the layers, with the main ones being layer complexity and performance, and accessibility. For instance, the RDDM is the most complex of the layers, as well as the best performing in terms of dwelling count allocation. However, this layer had to be specifically created for the study region, using a lengthy data collection, merging, and photointerpretation process. On the other end of the spectrum is the areal weighted census division layer, which simply took each CD and equally distributed the number of dwelling counts within. Notably, the areal weighted CD layer predicted a moderate number of exposed dwellings in the second evaluation, relative to the other 5 layers. Keep in mind however that with no clear relationship between each layer's allocation performance and its behavior when predicting exposed dwellings, the second evaluation requires further refinement.

Decision makers of course, are more focused on how these tools can perform for them, versus the variations in their construction. With this in mind, the RDDM would be the most accurate when trying to allocate dwelling counts across a region. This approach does however take more time and resources to carry out, meaning that the 12m GUF layer would be an ideal alternative for decision makers, as it had the second-best performance, and is already available at a global scale.

## 3.5 CONCLUSION

This chapter evaluated the Rural Dwelling Distribution Model created in chapter 1 and look at its ability to allocate dwellings across a given region, as well as its ability to predict exposed dwellings across several disaster scenarios. These results were then compared to other HSL approaches, to provide decision makers multiple options when needing measure and communicate risk in their communities.

The RDDM performed the best with regards to accurately allocating dwelling counts, while tending to predict the lowest number of exposed dwellings in the disaster scenarios. The 12m GUF layer on the other hand, predicted the second highest number of exposed dwellings, while still having the second-best performance in dwelling count allocation, behind the RDDM. This suggests that as we move from the regional to the site-specific level, there is not a clear relationship between HSL accuracy when allocating dwellings, and its behaviour when predicting exposed dwellings, and requires further investigation. Both the RDDM and the 12m GUF layer had the best and second-best allocation performances, respectively. This suggests that while the resource demanding and region-specific approach had the best performance, decision makers needing a more cost-effective alternative could use the 12m GUF layer, and still have comparable results.

## REFERENCES

Alberta Energy Regulator (2018). Pipelines [Vector Polyline layer]. Available from: https://www.aer.ca/providing-information/data-and-reports/maps-mapviewers-and-shapefiles.html

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal, 69(1–2), 103–117. https://doi.org/10.1007/s10708-007-9105-9

Center For International Earth Science Information Network (CIESIN) - Columbia University. (2016). Documentation for Gridded Population of the World, Version 4 (GPWv4). Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/h4d50jx4

Dobson, J., A. Bright, E., R. Coleman, P., C. Durfee, R., & A. Worley, B. (2000). LandScan: A Global Population Database for Estimating Populations at Risk. Photogrammetric Engineering and Remote Sensing, 66, 849–857.

Davies, J. B. (2016). Economic analysis of the costs of flooding. Canadian Water Resources Journal / Revue Canadienne Des Ressources Hydriques, 41(1–2), 204–219. https://doi.org/10.1080/07011784.2015.1055804

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., & Baptista, S. R. (2015). Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4. Papers in Applied Geography, 1(3), 226–234. https://doi.org/10.1080/23754931.2015.1014272

Elshorbagy, A., Bharath, R., Lakhanpal, A., Ceola, S., Montanari, A., & Lindenschmidt, K.-E. (2017). Topography- and nightlight-based national flood risk assessment in Canada. Hydrology and Earth System Sciences, 21(4), 2219–2232. https://doi.org/10.5194/hess-21-2219-2017

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., & Strano, E. (2017). Breaking new ground in mapping human settlements from space – The Global Urban Footprint. ISPRS Journal of Photogrammetry and Remote Sensing, 134, 30–42. https://doi.org/10.1016/j.isprsjprs.2017.10.012

Government of Canada (2016). 2016 Census of Canada (Provinces, Municipalities, Dissemination Areas). Available from: http://www5.statcan.gc.ca/cansim/home-accueil?lang=eng

Government of Canada (2020). Canada Energy Regulator – Provincial and Territorial Energy Profiles. Retrieved from: https://www.cer-rec.gc.ca/nrg/ntgrtd/mrkt/nrgsstmprfls/cda-eng.html

Hall, O., Stroh, E., & Payá, F. M. (2012). From Census to Grids: Comparing the Gridded Population of the World with Swedish Census Records. https://doi.org/10.2174/1874923201205010001

Klotz, M., Kemper, T., Geiß, C., Esch, T., & Taubenböck, H. (2016). How good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from Central Europe. Remote Sensing of Environment, 178, 191–212. https://doi.org/10.1016/j.rse.2016.03.001

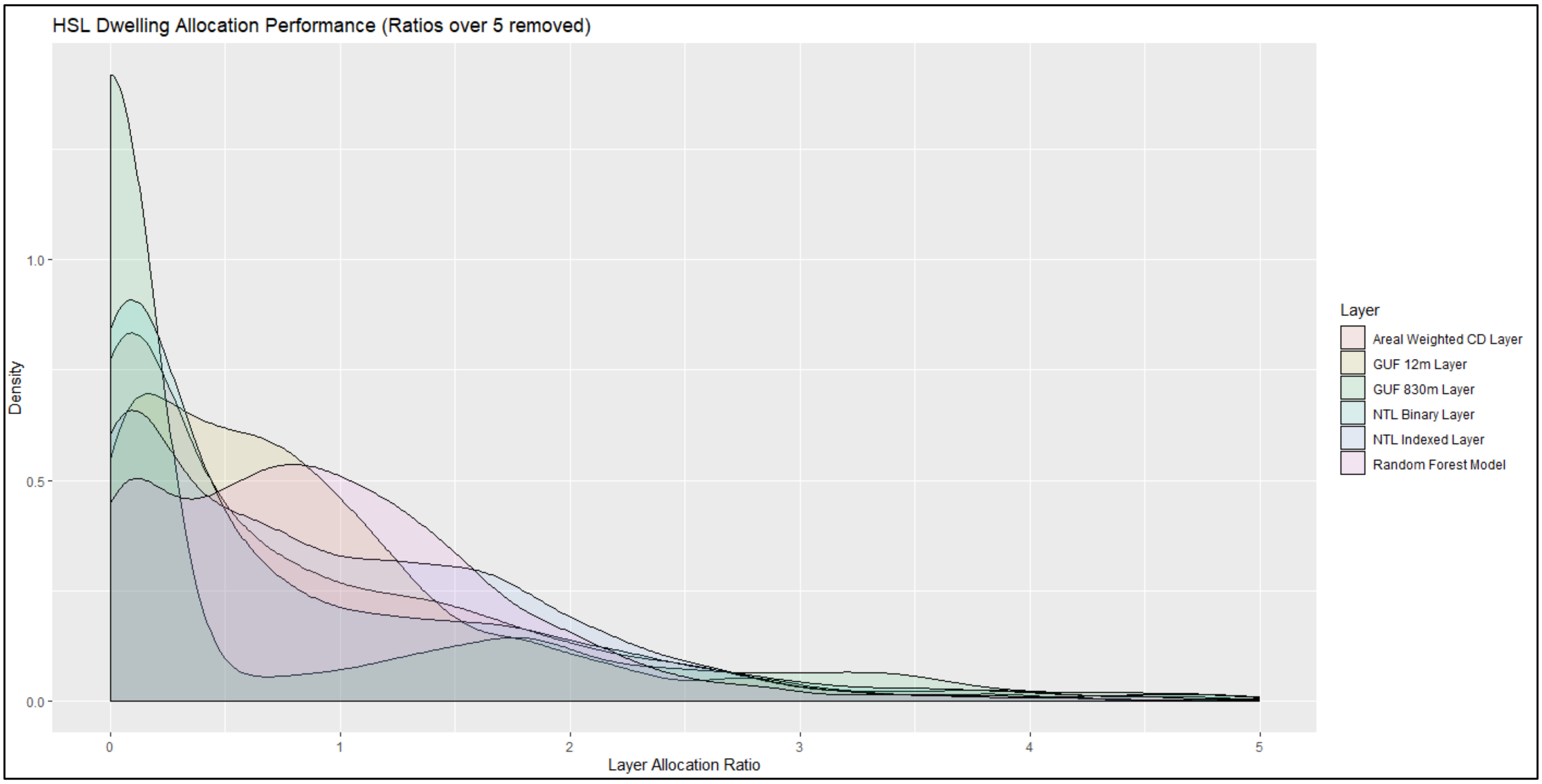
National Centers for Environmental Information (2016). VIIRS Day/Night Band Nighttime Lights [Raster Layer]. Available From: https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html

Province of Alberta (2017). Alberta ArcHydro Phase 2 [Vector Polyline layer]. Available from: https://www.alberta.ca/hydrological-data.aspx

Province of Alberta (2020). Municipal Government Act. Available from: https://www.qp.alberta.ca/1266.cfm?page=m26.cfm&leg_type=Acts&isbncln=9780779745739

Alberta Energy Regulator (2018). Pipelines [Vector Polyline layer]. Available from: https://www.aer.ca/providing-information/data-and-reports/maps-mapviewers-and-shapefiles.html

# Appendix A



*Appendix A.1: HSL Dwelling Allocation Density Plot*

# CHAPTER FOUR: CONCLUSION

## 4.1 INTRODUCTION

### 4.1.1 Summary of Major Findings

The aims of this research were to create a rural dwelling distribution model, or RDDM, for the province of Alberta using publicly accessible spatial data at the regional level. This model was trained using a random forest framework, and its ability to predict dwelling counts was evaluated against a test dataset. We then compared the RDDM to a univariate nighttime light layer, after converting both layers into binary surfaces.

Next, we evaluated the RDDM's ability to accurately allocate dwelling counts across a given region, as well as predict dwelling exposure in several disaster scenarios. These results were compared against several other human settlement layers, or HSLs, to determine whether creating a region-specific dwelling model is a recommended option for decision-makers, or if there are more feasible alternatives.

### 4.1.1 Chapter Two

Once created and trained, our region-specific dwelling model outperformed the univariate nighttime light layer in terms of accurately predicting dwelling presence. Overall, the RDDM tended to under predict dwellings in areas with already low dwelling counts, and when scaled up to the study area scale, we observe several clusters of predicted dwellings across the province, mainly seen near small communities and along the Edmonton-Calgary corridor (Chapter 2 Figure 13) . Furthermore, the RDDM had higher sensitivity to identifying dwellings in comparison to the nighttime light layer, which would be of more use to decision makers in rural areas, as minimizing the number of dwellings missed by an exposure model is crucial in a disaster scenario. In future research, removing outliers such as dense suburbs during the photointerpretation stage would be recommended, such as by creating a buffer around any excluded Towns, Urban Communities and Cities, to ensure that surrounding neighbourhoods are also not included.

### 4.1.2 Chapter Three

The rural dwelling distribution model had the best performance in terms of allocating dwelling counts across the provincial study area. Conversely, the 830m Global Urban Footprint (GUF) layer had the lowest number DA's that were accurately allocated dwellings, as well as having the highest amount of DA's that were either extremely under allocated, or extremely over allocated (Chapter 3 Table 8). With regards to all 6 layers, all but the RDDM and the 12m GUF layer tended to under allocate dwellings. These two layers also had the lowest amounts of error recorded, with 115.87 and 140.70 respectively (Chapter 3 Table 9).

With regards to the site-specific scenarios, the RDDM and the 830m GUF layers had the lowest and highest predicted amounts of dwellings, respectively (Chapter 3 Table 10). The 12m GUF layer had the second highest predicted amount of exposed dwellings as well as the second best allocation performance, suggesting that as we move from the region-specific to site-specific levels, there is not a clear relationship between a layer's ability to accurately allocate dwellings over a given region, and its ability to predict dwelling exposure. Overall, using a layer's dwelling count allocation performance was shown to be a valuable benchmark in determining its potential usefulness to decision makers, while the site-specific scenario evaluations require further refinement.

## 4.2 CONCLUSIONS

In this study, we created a region-specific rural dwelling distribution model for the province of Alberta and evaluated its ability to predict dwelling counts in comparison to a univariate nighttime light layer. Our results indicate that the RDDM outperformed the nighttime light layer in terms of both accuracy and sensitivity to identifying dwellings, suggesting that while it is more resource intensive, creating a multivariate region-specific dwelling model may perform better than a univariate nighttime light layer created at a global scale in terms of predicting dwellings.

We also determined that out of the 6 human settlement layers being measured, the RDDM had the best dwelling count allocation performance, as well as having the lowest amount of error, followed by the original 12m GUF layer in both categories.  These findings should be particularly valuable to decision makers, suggesting that creating a multivariate region-specific dwelling model would result in an effective dwelling distribution tool. As well, our findings suggest that there are also publicly available alternatives that can also be used to allocate dwelling count values, and while these layers may not be as accurate, they provide a cost-effective and suitable alternative.