Land Cover Classification of Fused Optical Radar Data Using Three Machine Learning Algorithms

Literature Review

Table of Contents

[Abstract 3](#_Toc95554400)

[Introduction 5](#_Toc95554401)

[Remote Sensing Classification Methodology 6](#_Toc95554402)

[Selection of remotely sensed data 7](#_Toc95554403)

[Selection of a classification system and training samples 7](#_Toc95554404)

[Data preprocessing and Feature extraction and selection 8](#_Toc95554405)

[Selection of a suitable classification method 8](#_Toc95554406)

[Post-classification processing 8](#_Toc95554407)

[Evaluation of classification performance 8](#_Toc95554408)

[Machine Learning Classifiers 9](#_Toc95554409)

[Support vector machine (SVM) 9](#_Toc95554410)

[Random forest (RF) 9](#_Toc95554411)

[K-nearest neighbours (k-NN) 10](#_Toc95554412)

[Naïve Bayes (NB) 10](#_Toc95554413)

[Study Design 10](#_Toc95554414)

[Methodology 11](#_Toc95554415)

[Project Approach 11](#_Toc95554416)

[Data Description 13](#_Toc95554417)

[References 14](#_Toc95554418)

[Appendix A 16](#_Toc95554419)

# Abstract

Remote sensing is the method of acquiring information about the physical characteristics of the earth’s surface from a distance, such as from satellites or aircraft, by measuring its reflected and emitted energy (NASA, 2021). Remote sensors can be either passive or active (NOAA, 2021). Passive sensors respond to external stimuli and record natural energy that is reflected or emitted from the Earth’s surface (NOAA, 2021). The most common source of reflected energy detected by passive sensors is sunlight (NOAA, 2021). Active sensors emit a stimulus to collect data about the Earth. For example, the active sensor LiDAR emits a laser-beam onto the Earth’s surface and measures the time it takes for the beam to reflect back to the sensor (NOAA, 2021). Remote sensors provide an abundance of data about the earth’s systems which facilitates data-informed decisions based on the current and future state of the globe (NASA, 2021). Remotely sensed information is used in military, intelligence, commercial, economic, planning and humanitarian applications, among others (Remote Sensing, 2022).

Remote sensing and data science communities have begun to merge due to several factors (Abdi, 2020). First, competitions held by companies such as Kaggle have demonstrated that high classification accuracies are possible using advanced machine learning algorithms (Abdi, 2020). Second, there is now an abundance of user-friendly open-source programming tools (Abdi, 2020). Third, a reduction in the cost of high-end computing power (Abdi, 2020). Fourth, and last, the availability of free remotely sensed data such as satellite data (Abdi, 2020).

One of the core applications of remotely sensed data is in land cover and land use classification. Land cover classification is the classification of the observed biophysical cover of the earth’s surface (NASA, 2021). Examples of land cover classes found on the earth’s surface include vegetation, urban infrastructure, water, rocks, and bare soil (Government of Canada, 2015). Land use classification is the classification of the purpose the land serves such as recreation, agriculture, or wildlife habitat (Government of Canada, 2015). Mapping land use/land cover (LULC) is important for global environmental monitoring studies, natural resources management, and landscape planning activities such as watershed management to name a few applications (Government of Canada, 2015).

The accuracy of LULC maps is dependent on several factors such as the training data used, the implementation of the classification, the classification algorithm, and the environmental heterogeneity of the area being classified (Hermosilla et al., 2022). Of these factors there is current interest in the application of advanced machine learning algorithms to LULC classification.

The research objectives of this project are to assess the performance of the K-Nearest Neighbours, Random Forest, and Support Vector Machine machine learning algorithms in the multi-class land cover classification of a fused optical-radar dataset (Kosravi, 2020). Using Python Pandas, Numpy, and Scikit-learn libraries each machine learning algorithm experiment will be applied to the same data set. Each model will be evaluated, and the machine learning algorithms compared to determine which one is best for our use case of classifying cropland from a fused optical-radar dataset for a relatively flat and environmentally homogenous area around Winnipeg, Manitoba.

# Literature Review

## Introduction

Remote sensing (RS) is the method of acquiring information about the earth’s surface from a distance, such as from aircraft or satellites, by measuring its reflected or emitted energy (NASA, 2021). One of the core applications of RS data is in land use/land cover (LULC) classification. Mapping of LULC using remote sensing is a cost-effective, spatially extensive, multi-temporal, and time-saving methodology (Talukdar et al., 2020).

Land cover classification is the classification of the observed biophysical cover of the earth’s surface (NASA, 2021). Land use is the classification of the purpose the land serves such as recreation, agriculture, or wildlife (Government of Canada, 2015). The change in land cover represents changes in land use categories and is caused by natural and man-made factors such as extreme weather events and urbanisation (Navin & Agilandeeswari, 2020).

Knowledge of LULC change is critical in various areas, such as environmental monitoring studies, natural resources management, and landscape planning activities such as watershed management (Government of Canada, 2015). The most efficient means of understanding landscape change is through the quantitative assessment and prediction of LULC dynamics (Talukdar et al., 2020).

Automatic classification of RS imagery has been a challenge for decades (van Leeuwen et al., 2020). Previously, the application of RS to LULC classification and its change was limited by the spatial and temporal resolution of the available sensors (van Leeuwen et al., 2020). With advances in the spatial and temporal resolution of satellite sensors, the input data for LULC change detection is in many cases available in abundance (van Leeuwen et al., 2020). It is now the situation that advanced machine learning (ML) algorithms can be used to process large amounts of data and produce accurate thematic maps in a timely manner over large areas (van Leeuwen et al., 2020).

ML techniques have been classified into two sub-types; supervised and unsupervised techniques (Talukdar et al., 2020). Supervised techniques are the most applicable to RS LULC classification and are characterized as using labelled datasets that train or “supervise” algorithms into classifying data or predicting outcomes accurately (Delua, 2021). Supervised ML techniques include algorithms such as support vector machine, random forest, k-nearest neighbours and naïve Bayes.

Over the last decade considerable attention has been spent on assessing the use of individual supervised ML algorithms as well as comparing the ML algorithms in RS applications (Talukdar et al., 2020). ML strengths when applied to the LULC mapping include the ability to handle datasets with high dimensionality and to map classes with very complex characteristics (Aguilera, 2020). In addition, the methods are nonparametric and do not make assumptions about data distribution and they can accept a variety of input predictor data (Aguilera, 2020). It has generally been found that ML algorithms produce higher accuracies than traditional statistical classifiers; especially when many predictor variables are involved (Maxwell et al. 2018).

## Study Area

The environmental characteristics of an area affects the performance of LULC classification. Studies have examined the application of ML classifiers to LULC classification in a variety of environmentally different study areas. Volke & Abarca-Del-Rio (2020) examined an earthquake and tsunami impacted coastal area of Chile. In Columbia, Aguilera (2020) assessed an open-pit mining area that contained forest, grass, bushes, urban area, water, and bare soil. The Caspian Hyrcanian mixed forests ecoregion of northern Iran was studied by Jamali (2019). This location was environmentally diverse and included plains, forest, rainforest and prairies (Jamali, 2019). Abdi (2020) assessed the performance of machine learning in a complex boreal landscape in south-central Sweden. Another biome specific study in the Brazilian tropical savanna (Cerrado) biome was conducted by Camargo et al. (2019). Lastly, flat temporary inundated areas in southern Hungary were examined by van Leeuwen et al. (2020).

## Remotely Sensed Imagery

The characteristics of the imagery used in a LULC classification is a factor

The purpose of this literature review is to examine studies that have researched the application of ML algorithms to LULC classification in the field of RS.

This review will inform the proposed project to assess the performance of the k-nearest neighbour, random forest, and support vector machine learning algorithms in the multi-class land cover classification of a fused optical-radar dataset (Kosravi, 2020).

# Remote Sensing Classification Methodology

RS applied to LULC classification is a complex process, requiring the consideration of many factors. The selection of remotely sensed data, the design of the classification procedure, and the quality of the classification result is influenced by the user’s need, scale of the study area, economic condition, and the analyst’s skills (Lu & Weng, 2007). The major steps involved in image classification are addressed in this section (Figure 1).

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure 1. Major steps involved in image classification (Lu & Weng, 2007).

## Selection of remotely sensed data

The selection of appropriate sensor data is the first step for a successful LULC classification project and understanding the strengths and weaknesses of different sensor data types is essential to this selection (Lu & Weng, 2007). The most important factors affecting the selection are the scale of the study area, image resolution and the user’s need (Lu and Weng, 2007). The user’s need determines the nature of the classification and the scale of the study area, which in turn affects the selection of the spatial resolution (Lu & Weng, 2007).

## Selection of a classification system and training samples

A land cover classification system is designed based on the user’s needs, spatial and spectral resolution of selected remotely sensed data, compatibility with previous studies, available classification and image processing algorithms and time constraints (Lu & Weng, 2007). It is important the spectral and spatial resolution of the imagery supports the separation of classes in the system and that the number of training samples are representative of the different classes (Lu & Weng, 2007).

## Data preprocessing and Feature extraction and selection

The quality of raw satellite data can be improved using preprocessing techniques such as atmospheric, radiometric, geometric, and topographic corrections (Navin & Agilandeeswari, 2020). Image enhancement techniques, such as principal component analysis (PCA), can be used to reduce the dimensionality of the satellite data (Navin & Agilandeeswari, 2020).

## Selection of a suitable classification method

When selecting a classification method many factors must be considered such as the different sources of data, the classification system, and the spatial resolution of the satellite imagery (Lu & Weng, 2007). The application of advanced ML algorithms to RS data is a burgeoning field. A selection of ML algorithms will be summarized in a later section of this document.

## Post-classification processing

The quality of a LULC classification can be improved with post-classification processing. Image classification is primarily based on the separation of spectral signatures. The complexity of biophysical environments can cause spectral confusion among LULC classes. A common method to resolve spectral confusion is to use ancillary data and expert knowledge to develop rule-based separation of LULC classes (Lu & Weng, 2007).

## Evaluation of classification performance

An accuracy assessment is the most important metric for validating a LULC classification (Navin & Agilandeeswari, 2020). The error matrix approach is the most prevalent accuracy assessment method (Lu & Weng, 2007). Given an error matrix the overall accuracy, omission error, commission error and Kapa coefficient can be calculated (Lu & Weng 2007). The Kappa coefficient is a measure of the overall statistical agreement of an error matrix (Lu & Weng, 2007). It is recognized as a good method to not only analyze a single error matrix but also compare the differences between error matrixes (Lu & Weng, 2007).

# Machine Learning Classifiers

This section’s purpose is to provide a conceptual description for a selection of ML algorithms used in the reviewed studies. Of special importance to the application of ML algorithms to remote sensing is the potential of overfitting. Overfitting occurs when a ML algorithm classifies the training data so precisely it is not able to generalize well to non-training data (Maxwell et al., 2018). Given the potential for overfitting, it is essential that when using ML algorithms in remote sensing new data not used in training the classifier be used to evaluate the accuracy of a LULC classification (Maxwell et al., 2018).

## Support vector machine (SVM)

SVM focuses on the training samples that are nearest in feature space to the ideal boundary between pairs of classes (Maxwell et al., 2018). These samples are called support vectors (Maxwell et al., 2018). Maximizing the separation between support vectors is the goal of SVM (Maxwell et al., 2018). Each possible combination of classes is processed by the SVM classifier such that the ideal single boundary between all classes is determined (Maxwell et al., 2018).

## Random forest (RF)

RF is an ensemble learning algorithm (Abdi, 2019). It uses multiple bootstrapped decision trees to overcome the weaknesses of a single decision tree (Maxwell et al., 2018). Bootstrapping means that the individual decision trees are parametrized using a random sample of observations with replacement from the training data (Abdi, 2019). Multicollinearity between decision trees is reduced by using bootstrapping (Abdi, 2019). Different groupings of the input variables are used to create the decision tree models (Abdi, 2019). The resulting classification is the unweighted majority vote for each class, averaged across decision trees (Abdi 2019).

## K-nearest neighbours (k-NN)

The k-NN algorithm assumes that similar things exist in close proximity (Harrison, 2018). Each unknown sample is compared against the training data and assigned to the most common class of the k training samples that are nearest in feature space (Maxwell et al., 2018). A low k will produce a complex decision boundary and a high k will result in greater generalization (Maxwell et al., 2018). k-NN is a computationally intensive algorithm (Maxwell et al. 2018).

## Naïve Bayes (NB)

The NB classifier uses the Bayesian theory pertaining to conditional probability and predictions of events, with strong independent assumptions (Camargo et al., 2019). The presence or absence of a given feature class is not dependent on the presence or absence of any other feature in the NB algorithm (Camargo et al., 2019). An advantage of the NB method is that it only requires a reduced amount of training data to estimate its parameters and thus can be trained very efficiently in a supervised learning framework (Camargo et al., 2019).

# Study Design

As part of the literature review the designs of various studies were examined. The designs can be divided into two parts: the RS processing of the satellite imagery and the application of the ML classifiers to produce the LULC classification. This section will examine the design of one typical study and will focus on the ML classifier methodology over the satellite preprocessing stage as the research project being considered attempts to apply that methodology

## Methodology

van Leeuwen et al. (2020) tested three different models to determine the best classification method, with each experiment being designed in the same way (Figure 2). The satellite data was preprocessed and then split into three sets; training, validation, and test data (van Leeuwen et al., 2020). For each model the optimal hyperparameters were determined and the model trained using the selected hyperparameters (van Leeuwen et al., 2020). The trained model was then tested using the test data and an accuracy assessment calculated using an error matrix and Cohen’s Kappa (van Leeuwen et al., 2020).

Diagram

Description automatically generated

Figure 2. Study methodology used by van Leeuwen et al., 2020.

# Project Approach

It has been found that there is interest in using advanced ML algorithms in RS. Numerous studies examined the use of ML algorithms for the LULC classification of RS data. Given the literature review it was found that the ML methodology outlined in Figure 3 was common among numerous studies. The methodology in Figure 3 will be applied to the proposed project.

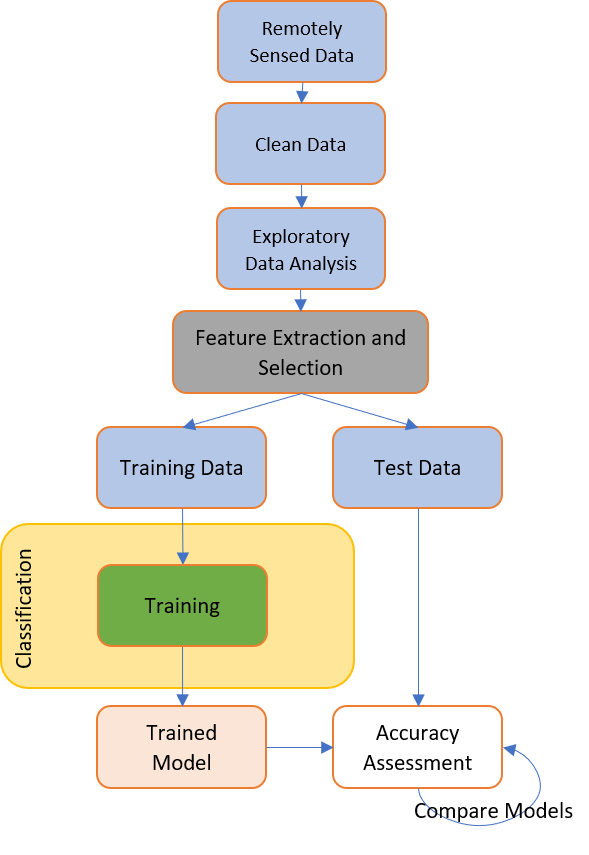


Figure 3. Proposed methodology for the project under consideration.

Given the interest in the application of ML algorithms to LULC classification in RS it is suggested that the project under consideration is worth conducting. The project gives the researcher the opportunity to practice using common ML techniques and apply them to a field of interest that is relevant.

# Data Description

The data to be used in this project is the “Crop mapping using fused optical-radar” dataset, created by Dr. Iman Khosravi (Khosravi, 2012). It is composed of fused, bi-temporal optical-radar data for cropland classification, in tabular form. The study area is an agricultural region near Winnipeg, Canada. The data is derived from images collected by RapidEye satellites (optical) and polarimetric radar data collected by Unmanned Aerial Vehicle Synthetic Aperture Radars (UAVSAR) on July 5th and July 14th, 2012. Seven crop classes exist in the dataset: 1 – corn; 2 – peas; 3 – canola; 4 – soy; 5 – oat; 6 – wheat; and 7 - broadleaf. The dataset has 175 attributes with 325,834 observations. The attributes are described in the Table 1 and a statistical summary of the features listed in Appendix A.

Table 1. Description of attributes in the cropland mapping dataset.

|  |  |
| --- | --- |
| **Features** | **Description** |
| Label | Crop type class |
| f1 to f49 | Polarimetric features on July 5, 2012 |
| f50 to f98 | Polarimetric features on July 14, 2012 |
| f99 to f136 | Optical features on July 5, 2012 |
| f137 to f174 | Optical features on July 14, 2012 |

# References

Abdi, A.M. (2020). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data*.* *GIScience & Remote Sensing, 57(1), 1-20.* <https://doi.org/10.1080/15481603.2019.1650447>

Camargo, F.F., Sano, E.E., Almeida, C.M., Mura, J.C., & Almeida, T. (2019). A comparative assessment of machine-learning techniques for land use and land cover classification of the Brazilian tropical savanna using ALOS-2/PALSAR-2 polarimetric images. *Remote Sensing, 11:1600*. <https://www.mdpi.com/2072-4292/11/13/1600>

Government of Canada. (2015, 11 20). *Land Cover and Land Use.* Land Cover/Biomass Mapping. <https://www.nrcan.gc.ca/maps-tools-and-publications/satellite-imagery-and-air-photos/tutorial-fundamentals-remote-sensing/educational-resources-applications/land-cover-biomass-mapping/land-cover-land-use/9373>

Hermosilla, T., Wulder, M. A., White, J. C., & Coops, N. C. (2022). Land cover classification in an era of big and open data: Optimizing localized implementation and training data selection to improve mapping outcomes. *Remote Sensing of Environment. 268:112780***.** <https://doi.org/10.1016/j.rse.2021.112780>

Delua, J. (2021, March 12). *Supervised vs. Unsupervised Learning: What’s the Difference?* IBM. <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>

Harrison, O. (2018, September 10). *Machine Learning Basics with the K-Nearest Neighbors Algorithm.* Towards Data Science. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

Khosravi, I. (2012). *Crop mapping using fused optical radar* [Data set]. UCI Machine Learning Repository. [https://archive.ics.uci.edu/ml/datasets/Crop+mapping+using+fused+optical-radar+data+set#](https://archive.ics.uci.edu/ml/datasets/Crop+mapping+using+fused+optical-radar+data+set)

Lu, D. & Weng, Q. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing. 28(5). 823-870.* <https://www.tandfonline.com/doi/full/10.1080/01431160600746456>

Maxwell, A.E., Warner, T.A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing. 39(9). 2784-2817.* <https://www.tandfonline.com/doi/full/10.1080/01431161.2018.1433343>

NASA. (2021, 12 6). *What is Remote Sensing?* Earth Data Open Access for Open Science. <https://earthdata.nasa.gov/learn/backgrounders/remote-sensing>

Navin, S.M., & Agilandeeswari, L. (2020). Comprehensive review on land use/land cover change classification in remote sensing. *Journal of Spectral Imaging. 9(a8).* <https://www.impopen.com/download.php?code=I09_a8>

NOAA. (2021, 2 26). *What is remote sensing?* Ocean Facts. <https://oceanservice.noaa.gov/facts/remotesensing.html>

Remote sensing. (2022, January 25). In *Wikipedia*. <https://en.wikipedia.org/wiki/Remote_sensing>

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations – A review. *Remote Sensing, 12:1135*. <https://www.mdpi.com/2072-4292/12/7/1135/htm>

van Leeuwen, B., Tobak, Z., & Kovacs, F. (2020). Comparison of different machine learning techniques for land use/land cover classification of medium resolution optical satellite imagery focusing on temporary inundated areas. **Journal of Environmental Geography, 13(1-2), 43-52.** <http://publicatio.bibl.u-szeged.hu/18958/1/10_2478-Voulme13_Issue_1_2_paper_5_Leeuwenetal.pdf>

# Appendix A

label f1 f2 f3 f4 f5 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 4.062 -15.144 -24.026 -15.400 -18.594 -14.493

std 1.605 3.504 4.054 3.268 3.712 3.189

min 1.000 -23.971 -34.308 -23.161 -27.245 -22.103

25% 3.000 -17.848 -27.119 -17.563 -21.449 -16.989

50% 4.000 -15.992 -25.064 -16.164 -19.588 -15.099

75% 6.000 -11.786 -20.387 -13.427 -15.650 -11.735

max 7.000 2.536 -7.589 1.104 0.775 -0.887

f6 f7 f8 f9 f10 f11 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean -18.427 0.256 -8.882 -8.626 -0.166 4.100

std 3.676 1.682 1.204 2.003 0.596 1.500

min -26.897 -5.583 -16.389 -16.558 -2.717 -4.282

25% -21.450 -1.038 -9.659 -10.070 -0.526 3.141

50% -19.282 0.309 -8.916 -8.733 -0.175 4.079

75% -15.514 1.558 -8.190 -7.158 0.172 5.090

max 0.952 7.030 -1.175 -0.925 3.785 11.212

f12 f13 f14 f15 f16 f17 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 3.934 0.480 0.064 0.456 0.218 0.556

std 1.539 0.084 0.020 0.093 0.042 0.082

min -4.498 0.209 0.016 0.162 0.065 0.154

25% 2.933 0.415 0.049 0.383 0.190 0.503

50% 3.958 0.483 0.061 0.450 0.217 0.556

75% 4.880 0.548 0.074 0.529 0.245 0.612

max 10.452 0.819 0.272 0.756 0.437 0.854

f18 f19 f20 f21 f22 f23 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.227 -1.517 -1.609 -1.181 -1.593 -0.309

std 0.045 0.644 0.622 0.520 0.668 0.155

min 0.076 -6.793 -7.757 -7.643 -7.677 -2.051

25% 0.198 -1.836 -1.917 -1.436 -1.957 -0.386

50% 0.224 -1.424 -1.496 -1.083 -1.510 -0.277

75% 0.256 -1.084 -1.178 -0.837 -1.143 -0.203

max 0.434 -0.210 -0.374 -0.124 -0.231 -0.028

f24 f25 f26 f27 f28 f29 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean -1.487 0.058 0.025 0.005 0.715 0.703

std 0.635 0.054 0.020 0.005 0.067 0.105

min -7.384 0.006 0.003 0.000 0.307 0.145

25% -1.786 0.023 0.011 0.002 0.679 0.637

50% -1.377 0.035 0.016 0.003 0.718 0.710

75% -1.059 0.078 0.036 0.007 0.762 0.778

max -0.200 2.445 0.763 0.113 0.929 0.969

f30 f31 f32 f33 f34 f35 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 41.669 0.500 0.215 0.202 0.082 0.083

std 5.776 0.076 0.088 0.060 0.030 0.038

min 17.538 0.102 0.020 0.024 0.011 0.007

25% 37.920 0.447 0.157 0.165 0.059 0.056

50% 41.985 0.506 0.199 0.202 0.078 0.076

75% 45.504 0.563 0.260 0.243 0.100 0.099

max 73.211 0.649 0.665 0.522 0.400 0.357

f36 f37 f38 f39 f40 f41 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.210 -14.493 -16.191 -24.026 0.143 0.128

std 0.084 3.189 3.664 4.054 0.054 0.060

min 0.022 -22.103 -24.520 -34.308 0.057 0.045

25% 0.151 -16.989 -19.112 -27.119 0.101 0.082

50% 0.194 -15.099 -17.046 -25.064 0.125 0.104

75% 0.248 -11.735 -13.388 -20.387 0.183 0.164

max 0.690 -0.887 3.555 -7.589 0.625 1.055

f42 f43 f44 f45 f46 f47 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.006 0.035 0.029 0.025 0.037 0.029

std 0.005 0.028 0.038 0.024 0.027 0.035

min 0.000 0.006 0.006 0.006 0.006 0.006

25% 0.002 0.016 0.009 0.008 0.017 0.010

50% 0.005 0.025 0.015 0.013 0.027 0.016

75% 0.009 0.047 0.032 0.036 0.051 0.036

max 0.064 0.423 1.987 0.640 0.484 1.971

f48 f49 f50 f51 f52 f53 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.006 0.021 -17.033 -26.718 -18.747 -21.423

std 0.001 0.022 3.568 4.046 3.181 3.505

min 0.000 0.006 -24.812 -35.604 -26.894 -28.954

25% 0.006 0.006 -19.843 -30.148 -20.944 -24.245

50% 0.006 0.010 -18.036 -27.312 -19.225 -21.941

75% 0.006 0.030 -14.484 -23.788 -17.326 -19.261

max 0.057 0.551 2.982 -8.711 -2.500 -1.481

f54 f55 f56 f57 f58 f59 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean -16.771 -21.236 1.714 -9.685 -7.970 -0.187

std 3.250 3.613 1.744 1.599 1.885 0.575

min -24.015 -28.315 -4.367 -18.304 -14.727 -4.721

25% -19.258 -24.232 0.240 -10.706 -9.067 -0.568

50% -17.750 -21.583 1.549 -9.749 -7.860 -0.217

75% -14.935 -19.010 3.228 -8.629 -6.575 0.170

max 0.444 -1.133 9.722 -1.760 0.779 4.729

f60 f61 f62 f63 f64 f65 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 4.652 4.464 0.557 0.062 0.381 0.202

std 1.070 1.053 0.089 0.020 0.089 0.032

min -1.471 -1.516 0.243 0.011 0.094 0.078

25% 3.993 3.738 0.481 0.048 0.300 0.183

50% 4.601 4.418 0.555 0.062 0.387 0.202

75% 5.293 5.126 0.630 0.074 0.454 0.220

max 10.230 9.245 0.885 0.279 0.681 0.400

f66 f67 f68 f69 f70 f71 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.586 0.211 -2.084 -1.259 -1.976 -2.266

std 0.056 0.031 0.681 0.370 0.627 0.690

min 0.262 0.098 -8.565 -5.417 -8.266 -8.127

25% 0.549 0.190 -2.467 -1.474 -2.340 -2.651

50% 0.584 0.211 -2.011 -1.208 -1.905 -2.163

75% 0.622 0.232 -1.593 -1.004 -1.510 -1.744

max 0.824 0.395 -0.573 -0.248 -0.581 -0.558

f72 f73 f74 f75 f76 f77 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean -1.884 -1.478 0.033 0.015 0.003 0.719

std 0.599 0.668 0.035 0.015 0.004 0.057

min -7.913 -7.232 0.005 0.002 0.000 0.300

25% -2.218 -1.877 0.013 0.006 0.001 0.686

50% -1.788 -1.385 0.019 0.009 0.002 0.723

75% -1.461 -0.928 0.038 0.015 0.004 0.757

max -0.591 -0.144 2.047 0.510 0.124 0.969

f78 f79 f80 f81 f82 f83 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.660 41.743 0.474 0.246 0.187 0.094

std 0.102 6.021 0.075 0.081 0.052 0.031

min 0.098 16.952 0.058 0.036 0.005 0.026

25% 0.590 37.074 0.412 0.190 0.152 0.069

50% 0.667 42.301 0.477 0.237 0.181 0.089

75% 0.734 46.773 0.536 0.296 0.220 0.117

max 0.921 62.363 0.641 0.812 0.501 0.365

f84 f85 f86 f87 f88 f89 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.090 0.232 -16.771 -19.029 -26.718 0.111

std 0.032 0.074 3.250 3.513 4.046 0.048

min 0.009 0.029 -24.015 -26.255 -35.604 0.047

25% 0.067 0.179 -19.258 -21.923 -30.148 0.077

50% 0.088 0.230 -17.750 -19.450 -27.312 0.092

75% 0.107 0.273 -14.935 -16.912 -23.788 0.127

max 0.525 0.932 0.444 1.589 -8.711 0.719

f90 f91 f92 f93 f94 f95 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.091 0.005 0.025 0.013 0.014 0.025

std 0.042 0.005 0.025 0.017 0.015 0.025

min 0.036 0.000 0.004 0.004 0.004 0.004

25% 0.060 0.002 0.010 0.005 0.004 0.011

50% 0.079 0.004 0.014 0.008 0.008 0.015

75% 0.108 0.007 0.029 0.013 0.017 0.029

max 0.809 0.076 0.727 1.648 0.512 0.709

f96 f97 f98 f99 f100 f101 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.014 0.004 0.011 0.005 0.053 0.037

std 0.017 0.001 0.013 0.007 0.025 0.023

min 0.004 0.000 0.004 0.001 0.001 0.001

25% 0.005 0.004 0.004 0.001 0.034 0.017

50% 0.008 0.004 0.006 0.001 0.044 0.032

75% 0.015 0.004 0.014 0.009 0.067 0.054

max 1.590 0.055 0.394 0.279 0.341 0.321

f102 f103 f104 f105 f106 f107 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.127 0.417 0.826 18.178 1.702 0.586

std 0.039 0.119 0.115 20.225 0.843 0.154

min 0.001 0.001 -0.018 0.965 0.071 -0.029

25% 0.098 0.322 0.757 7.233 1.234 0.470

50% 0.111 0.423 0.846 12.000 1.500 0.634

75% 0.148 0.497 0.924 25.381 2.000 0.702

max 0.342 0.752 0.996 556.000 34.000 1.020

f108 f109 f110 f111 f112 f113 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.708 0.584 0.219 0.769 0.630 0.523

std 0.170 0.128 0.160 0.095 0.199 0.110

min -0.188 -0.013 -0.867 -0.127 -0.034 -0.750

25% 0.590 0.495 0.105 0.702 0.478 0.452

50% 0.729 0.632 0.200 0.771 0.682 0.520

75% 0.858 0.678 0.333 0.852 0.778 0.622

max 0.996 0.873 0.943 0.996 1.187 0.758

f114 f115 f116 f117 f118 f119 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 3.411 -0.433 25.317 0.581 0.065 6.025

std 0.991 0.085 8.941 0.152 0.157 1.866

min 0.143 -0.976 -0.860 -0.042 -5.889 -2.540

25% 2.648 -0.496 17.760 0.457 0.015 4.800

50% 3.168 -0.431 27.730 0.586 0.087 5.620

75% 4.293 -0.366 32.200 0.712 0.130 7.080

max 7.273 0.116 51.500 0.982 0.279 12.340

f120 f121 f122 f123 f124 f125 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 4.694 25.880 1.263 0.646 2.399 0.922

std 3.730 7.339 3.450 0.182 5.784 0.813

min 0.920 1.556 0.000 0.012 0.000 0.000

25% 2.680 20.111 0.222 0.533 0.444 0.444

50% 3.833 25.889 0.444 0.667 0.889 0.667

75% 5.941 30.556 0.988 0.778 1.889 1.111

max 111.000 45.000 123.560 1.000 169.780 11.000

f126 f127 f128 f129 f130 f131 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 1.414 0.301 0.279 3.904 0.139 0.881

std 0.454 0.167 0.404 1.932 0.487 0.124

min -0.000 0.111 -1.000 0.000 0.000 0.018

25% 1.149 0.185 0.000 2.000 0.000 0.778

50% 1.465 0.259 0.335 4.000 0.099 0.889

75% 1.735 0.358 0.601 5.556 0.222 1.000

max 2.197 1.000 1.000 20.000 125.730 1.000

f132 f133 f134 f135 f136 f137 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.296 0.246 0.579 0.670 0.536 6155.009

std 0.905 0.286 0.524 0.287 0.501 245.682

min 0.000 0.000 -0.000 0.111 -1.000 5495.000

25% 0.000 0.000 -0.000 0.407 0.158 5959.000

50% 0.222 0.222 0.530 0.654 0.632 6098.000

75% 0.444 0.444 1.003 1.000 1.000 6368.000

max 290.330 14.111 2.197 1.000 1.000 14311.000

f138 f139 f140 f141 f142 f143 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 5429.190 3224.351 4651.873 10335.964 0.513 3.302

std 582.647 572.939 746.884 2262.584 0.117 0.869

min 4310.000 2219.000 2462.000 3050.000 -0.101 0.817

25% 5011.000 2847.000 4129.000 8827.000 0.477 2.826

50% 5214.000 3091.000 4415.000 9872.000 0.529 3.247

75% 5800.000 3404.000 5254.000 11906.000 0.586 3.827

max 12687.000 10768.000 8734.000 18281.000 0.750 6.994

f144 f145 f146 f147 f148 f149 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 1.709 -1.166 0.949 0.769 0.258 0.300

std 0.180 1.970 0.187 0.176 0.054 0.090

min 1.030 -502.770 0.003 -0.151 0.015 -0.273

25% 1.621 -1.343 0.895 0.716 0.237 0.263

50% 1.715 -0.977 0.985 0.794 0.263 0.309

75% 1.831 -0.799 1.063 0.879 0.294 0.356

max 2.293 912.600 1.980 1.125 0.393 0.530

f150 f151 f152 f153 f154 f155 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.811 0.372 2.235 0.372 519341.312 0.181

std 0.127 0.085 0.400 0.085 169918.820 0.067

min -0.035 -0.020 0.961 -0.020 6390.000 -0.104

25% 0.798 0.339 2.026 0.339 419620.000 0.148

50% 0.837 0.379 2.220 0.379 511300.000 0.175

75% 0.877 0.419 2.441 0.419 631180.000 0.219

max 0.956 0.585 3.819 0.585 1105900.000 0.416

f156 f157 f158 f159 f160 f161 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 4939.284 173844.894 1.459 44.453 0.480 0.756

std 1758.171 54569.523 0.203 5.066 1.264 0.165

min -4178.200 -45280.000 0.811 27.000 0.000 0.028

25% 3933.000 143760.000 1.347 40.889 0.099 0.667

50% 4624.100 163940.000 1.424 45.778 0.222 0.778

75% 6116.150 210115.000 1.562 47.889 0.444 0.889

max 10862.000 365400.000 2.422 60.889 93.802 1.000

f162 f163 f164 f165 f166 f167 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.930 0.554 1.075 0.434 0.329 2.171

std 2.121 0.491 0.542 0.246 0.442 1.285

min 0.000 0.000 -0.000 0.111 -1.000 0.000

25% 0.222 0.222 0.687 0.259 0.000 1.000

50% 0.444 0.444 1.149 0.358 0.357 2.000

75% 0.889 0.667 1.465 0.506 0.661 3.000

max 134.890 9.333 2.197 1.000 1.000 12.444

f168 f169 f170 f171 f172 f173 \

count 325834.000 325834.000 325834.000 325834.000 325834.000 325834.000

mean 0.080 0.920 0.174 0.163 0.403 0.765

std 0.183 0.106 0.374 0.222 0.480 0.274

min 0.000 0.106 0.000 0.000 -0.000 0.111

25% 0.000 0.833 0.000 0.000 -0.000 0.506

50% 0.000 1.000 0.000 0.000 -0.000 1.000

75% 0.173 1.000 0.333 0.333 0.849 1.000

max 25.951 1.000 66.667 6.667 2.197 1.000

f174

count 325834.000

mean 0.668

std 0.471

min -1.000

25% 0.357

50% 1.000

75% 1.000

max 1.000