[AM-08-094] Machine Learning Approaches

Abstract

Machine learning approaches are increasingly used across numerous applications in order to learn from data and generate new knowledge discoveries, advance scientific studies and support automated decision making. In this knowledge entry, the fundamentals of Machine Learning (ML) are introduced, focusing on how feature spaces, models and algorithms are being developed and applied in geospatial studies. An example of a ML workflow for supervised/unsupervised learning is also introduced. The main challenges in ML approaches and our vision for future work are discussed at the end.

Keywords: geospatial data science, inference, machine learning, prediction, statistical learning

Author & citation

Wachowicz, M. and Gao, S. (2020). Machine Learning Approaches. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2020 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2020.2.5.

Explanation

1. Fundamentals of Machine Learning (ML)
2. How to select a ML Algorithm
3. Workflow and Applications
4. Challenges and a Vision for the Future

1. Fundamentals of Machine Learning

Field of study that gives computers the ability to learn without being explicitly programmed. (definition of machine learning, Arthur Samuel, 1959)

Machine Learning (ML) was originally coined by Arthur Samuel in the late fifties. It is considered as a field of Artificial Intelligence (AI) based on the concept that machines can learn from data and make decisions with minimal human intervention (Samuel 1967; Michie 1968). High performance computing is required for implementing ML models because they include feature spaces (also known as data spaces) that consist of vast amounts of data, having nominal, ordinal, interval and ratio measurement scales (Stevens 1946; Hand 2016). Building feature spaces using nominal scales is usually limited since they are the lowest level of measurement in which the only empirical measurement is that objects have different values of an attribute. In this case, constructing attributes of interest over time plays an important role in the performance and accuracy of a ML model. In contrast, using ordinal measurement scales, the order is meaningful but the actual values are not. Building a feature space involves establishing a mapping from objects and their relationships. In this case, it makes no sense to concatenate objects to yield a new object to be part of a feature space. Taking into account the interval and ratio scales, the measurements require that the difference between two values is meaningful. In this case, both the order relationship and

the concatenation of differences between objects must be reflected in the relationships between measurements of a feature space. Exploring interval and ratio measurement scales in building feature spaces for a ML model is a complex task because of the explicit mapping from an empirical structure in the real-world to a numerical representation in a feature space.

In a feature space, we distinguish two types of measurements. The outcome measurement (also known as a dependent variable or a feature label) that we wish to obtain based on analyzing a set of feature measurements (also known as independent variables). A training set of data is used to observe both the outcome and feature measurements for a-priori known objects, and fit an ML model that will enable us later to predict the outcome measurement for a new object. A validation data set is used to estimate the generalization error in order to choose a reliable ML model that accurately predicts an outcome measurement. After having chosen a final ML model, estimating its generalization error on new objects is carried out using a test data set. The assessment of a ML model depends on its prediction capability on independent feature measurements which have not been used in the training data set.

It is a multifaceted task to choose the appropriate number of measurements in each training, validation, and testing data sets. A typical split is 60% for training, 20% for validation, and 20% for testing. It is also challenging to devise a general rule on how much training data is enough since it will depend on the complexity of the ML model being used. Figure 1 shows what might happen as the number of training samples is increased when using a ML model. The error on the training set actually increases, while the generalization error, i.e. the error on an unseen test sample, decreases. The two converge to an asymptote. Naturally, ML models with different complexity have asymptotes at different error values. In Figure 1, model 1 converges to a higher error rate than model 2 because model 1 is not complex enough to capture the structure of the training data set. However, if the amount of training data available is less than a certain threshold, then the less complex model 1 wins. This regime is important because often the amount of training data is fixed — it is just not possible to obtain any more training data.

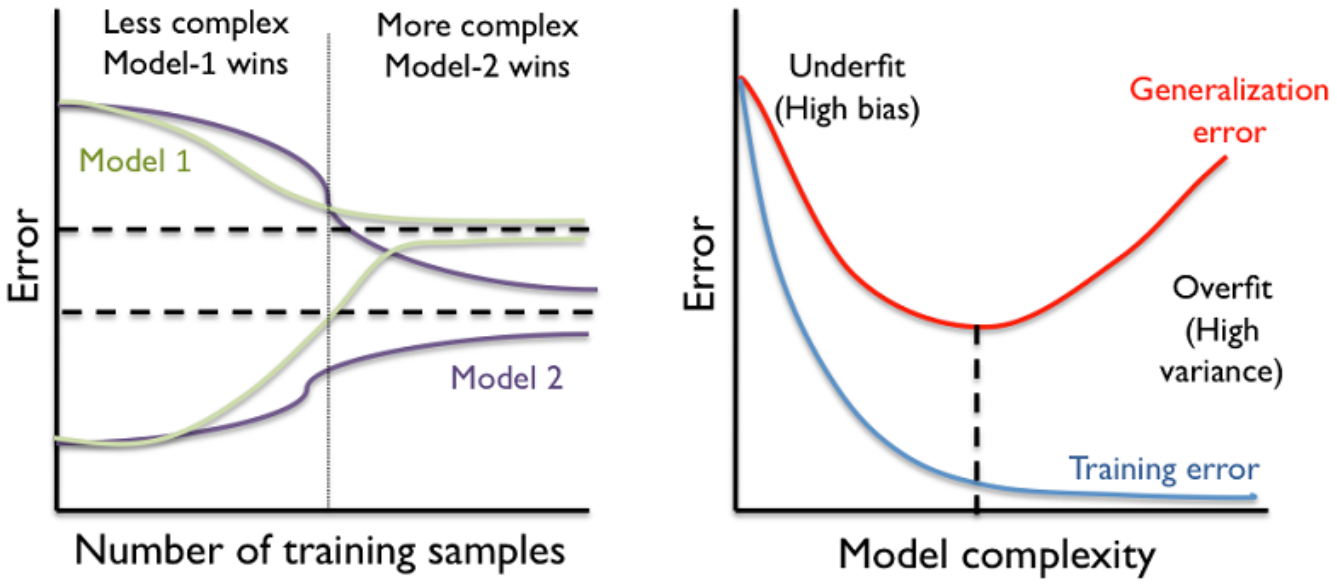


Figure 1. ML Model Complexity according to the number of training data sets and bias/variance. Image source: authors.

Figure 1 also shows the change in training error and generalization error as the model complexity increases, and the size of the training set is held constant. A model is said to underfit the training data when the model performs poorly on the training data and fails to learn the relationship between features and the target outputs because of high training error (high bias). As the model complexity increases, the training set error decreases monotonically, but the generalization error first falls and then increases. This occurs because by choosing a progressively complex model, at some point, the model begins to overfit the training data. That is, given the larger number of degrees of freedom in a more complex model, the model begins to adapt to the noise present in the dataset. This has a negative impact on generalization error. Thus, an overfitting model performs well on the training data but does not perform well on the testing data with high variance. One way to address this issue is to regularize the ML model — i.e. somehow constrain the model parameters, which implicitly reduces the degrees of freedom. If the learning problem can be framed as an optimization problem (e.g. find minimum mean square error), then one way to regularize is to alter the objective function by adding on a term involving the model parameters. Regularization is a flexible means to control the model's capacity in order to avoid overfitting or underfitting.

If application of an ML model shows that training and generalization error rates have leveled out, close to their asymptotic value, then adding additional training samples is not going to help. The only way to improve generalization error would be to use a different, more complex ML model. On the other hand, if all the available training samples have been used, and increasing the model's complexity seems to be increasing the generalization error, then this points to overfitting. In this case, generalization error is usually measured by keeping aside part of the training data (for example 30%) and using only the remaining data for training. A more common technique, termed k-fold cross validation, is to use a smaller portion (for example 5% or 10%) as test data, but repeat k-times with random portions kept aside as test data.

Many learning processes have been developed for ML models. Traditionally supervised and unsupervised learning models have been widely used in machine learning. In supervised learning models, the presence of the outcome measurements is used to guide the learning process since we have a direct feedback to predict the outcome/future. Two types of outcomes are predicted in supervised ML models: continuous numeric values (regression) and discrete categorical values (classification). In contrast, in unsupervised learning we have no outcome measurements as well as no feedback. The aim of an unsupervised learning model is actually to find hidden structure and describe how the feature measurements are organized or clustered. Clustering models learn to gather a set of objects such that objects in the same group are more similar to each other than to those objects in other groups. Supervised and unsupervised models are developed under a common assumption: the training and test data are drawn from the same feature space and the same probabilistic distribution.

Additional learning processes have been proposed in the literature. One example is semi-supervised learning which aims to understand how combining a small amount of labeled

data with a large amount of unlabeled data may change the data mining and ML behavior (Zhu & Goldberg 2009; Vatsavai et al. 2005). Another example includes transfer learning which is a process for training an ML model trained in one time period (the source domain) into a new time period (the target domain). In many applications the outcome measurement obtained in one time period may not follow a similar probabilistic distribution in a later time period. Transfer learning is usually selected for a ML model when training data can become regularly outdated. This is particular the case for data generated by the Internet of Things (IoT). The IoT devices are usually equipped with different types of sensors including accelerometers, gyroscopes, GPS, light, noise, motion, microphones, and cameras that seamlessly interact with the environment and sense feedback which will guide the learning process (Atzori et al. 2010).

Another example is ensemble learning that uses a set of models, each of them obtained by applying a learning process to a given problem, either in classification or regression problems. This set of ensembles (models) is aggregated in linearly weighted ensembles or majority voting to obtain a combined ML model that outperforms every single ensemble in it. The advantage of combined ML models with respect to single models has been reported in terms of increased robustness and accuracy (Mendes-Moreira et al 2012). Bias-variance decomposition and strength correlation usually explain why ensemble methods work in a variety of applications.

Finally, reinforcement learning differs from the previous learning processes in a fundamental way because there is no feature space with outcome and feature measurements. Instead, the learning process is “represented by an agent connected to its environment via perception and action. On each step of the learning process an agent receives as input some measurements of the current state of an environment; the agent then chooses an action to generate an output. This action changes the state of the environment, and the value of this state transition is communicated to the agent through a scalar reinforcement signal. The agent's behavior should choose actions that tend to increase the long-run sum of values of the reinforcement signal. It can learn to do this overtime by systematic trial and error, guided by a wide variety of algorithms” (Kaelbling et al. 1996, p. 238). However, reproducing the learning process is a rarely straightforward task, and previous research work describes a wide range of outcomes for the same ML models.

2. How to Select an ML Algorithm

There is an ML algorithm for each learning process. The objective of supervised ML algorithms can be described as the same as a statistical method. They both aim at improving accuracy by minimizing some function, typically the sum of squared errors (in regression problems). Their difference lies in how such a minimization is carried out using ML algorithms. Non-linear methods are used in ML algorithms meanwhile statistical ones use linear methods (Hastie et al. 2009). Decision trees are an example of a low-bias algorithm, whereas linear regression is an example of a high-bias algorithm. The k-Nearest Neighbors (KNN) algorithm is an example of a high-variance algorithm, whereas Linear Discriminant Analysis is an example of a low-variance algorithm. The parameterization of ML algorithms is often a battle to balance out bias and variance because increasing the bias will decrease the variance and increasing the variance will decrease the bias. In general, ML

algorithms may take a long time to generate the outputs and high-performance computing is needed to train large volumes of training data sets.

The first step to select an ML algorithm is to realize that you will be working on a multidimensional feature space that will be used to learn from data. The workflow to select an ML algorithm can be described as follows:

- Frame your question in the context of a hypothetical function (f) that the ML algorithm aims to learn. Given some input variables (Input) the function answers the question as to what is the predicted output variable (Output). The inputs and outputs can be referred to as variables or vectors.

Output = f(Input)

- In ML algorithms, a hyperparameter is a parameter whose value is set manually by the data scientist before the learning process begins. Given these hyperparameters, the ML algorithm learns the values of the other parameters of a ML algorithm using the training data. In other words, these parameters are estimated from data and they are often saved as part of a ML model.
- Algorithms that simplify the function to a known form are called parametric machine learning algorithms. Some examples of parametric ML algorithms are Linear Regression and Linear Discriminant Analysis (LDA). Two steps are involved: (1) select a form of function, (2) learn the coefficients for the function from the training data. Generally, parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithm's bias.
- Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric ML algorithms. By not making assumptions, they are free to learn any functional form from the training data. They are often more flexible, achieve better accuracy but require a lot more data and training time. Examples of nonparametric ML algorithms include Support Vector Machines, Deep Learning, Neural Networks and Decision Trees.

There is currently a vast number of ML algorithms available in the literature. For classification problems alone, Fernandez-Delgado et al. (2014) evaluated 179 classifiers from 17 families of ML algorithms searching to answer if we actually need hundreds of classifiers to solve real-world classification problems. The proliferation of classifiers is usually due to "each time we find a new classifier or family of classifiers from areas outside our domain of expertise, we ask ourselves whether that classifier will work better than the ones that we use routinely (Fernandez-Delgado et al., 2014 p. 3134). Therefore, Table 1 aims to provide an overview of the most used families of ML algorithms (Bishop 2006; Pedregosa et al. 2011).

Table 1. Most Common Families of Machine Learning Algorithms

Family	Main Characteristics
Linear Regression	• Linear regression is perhaps one of the most well-known and well-understood method in statistics and machine learning. • It is a fast and simple technique and a good first algorithm to try.

Family	Main Characteristics
Logistic Regression	• It is NOT a regression method. Unlike linear regression, the prediction for the output is transformed using a nonlinear function called the logistic function. • It is another technique borrowed by machine learning from the field of statistics. • It is the go-to method for binary classification problems (problems with two class values).
Linear Discriminant Analysis (LDA)	• The technique assumes that the data has a Gaussian distribution (bell curve), therefore it is important to remove outliers from your data beforehand. • If you have more than two classes then the LDA algorithm is the preferred linear classification technique. • It is a simple and powerful method for classification predictive learning problems.
Decision Trees	• Decision tree use tree-like structure to explicitly represent the process of decision making. • Trees are fast to learn and very fast for making predictions. They are also often accurate for a broad range of problems (regression and classification) and do not require any special preparation for your data. • Decision trees have a high variance but a low-bias and can yield more accurate predictions when used in an ensemble learning process.
Random Forest	• Random forest is an improved model over bagged decision trees and changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation. • Multiple samples of your training data are taken then models are constructed for each data sample. When you need to make a prediction for a new measurement, each model makes a prediction and the predictions are averaged to give a better estimate of the true output value. • If you get good results with an algorithm with high variance (like decision trees), you can often get better results by bagging that algorithm.
Bayesian	• Naive Bayes classifier is a conditional probability model and applies Bayes' theorem for mapping a prior to a posterior. • Naive Bayes is called naive because it assumes that each input variable is independent. This is a strong assumption and unrealistic for real-world data, nevertheless, the technique is very effective on a large range of complex classification problems.
Support Vector Machine (SVM)	• The learning process finds the coefficients that results in the best separation of the classes by a hyperplane. • In practice, an optimization algorithm is used to find the values for the coefficients that maximizes the margin. • SVM might be one of the most robust out-of-the-box classifiers and worth trying on your dataset.
K Nearest Neighbors (KNN)	• KNN is a non-parametric supervised learning method used for classification and regression based on k closest training examples in the feature space. Note that KNN is different from the unsupervised learning method k-means clustering. • The challenge is in how to determine the similarity between the measurements. The simplest technique if your measurements are all of the same measurement scale is to use the Euclidean distance. • KNN can require a lot of memory to store all of the data, but only learns when a prediction is needed. • The concept of distance or closeness can be an issue in very high dimensions (lots of feature measurements) which can negatively affect the performance of the algorithm on your problem. This is called the curse of dimensionality. You should only use those input variables that are most relevant to predicting the output variable.
Deep Neural Networks	• A family of machine learning techniques with layers of neurons, where many layers of information processing stages in hierarchical neural net architectures, are exploited for unsupervised feature learning and for supervised pattern classification and regression.

In general, spatial data can be used in feature spaces of any family of ML algorithms, but there is a caveat: it is not a straightforward decision about how geographical locations, spatial constraints, and geometrical and topological relationships among objects can be represented within a feature space. Point coordinates or more complex structures such as density information or graph structures have been used to represent the spatial dimension in a feature space. The main goal is that a geographical space is embedded in a feature

space of a ML model. Some examples include K-means spatial clustering (Jain 2010), density-based spatial clustering (DBSCAN) algorithm, grid-based spatial clustering (GCHL), anisotropic density-based clustering (ADCN), and hierarchical spatial clustering algorithm (HDBSCAN) (Ester et al. 1996; Pilevar & Sukumar 2005; Campello et al. 2013; Mai et al. 2018).

In addition to the application of traditional machine learning techniques to spatial data, latent representation learning and spatially explicit machine learning techniques by considering spatial constraints and relationships have attracted much attention in geospatial data science (Bengio et al. 2013; Aydin et al. 2018; Singleton & Arribas-Bel 2019; Yan et al. 2019; Janowicz et al. 2020). Researchers have utilized representation learning for latent geospatial feature representation, such as place type representation (Yan et al. 2017; Zhang et al. 2017), points of interest (POI) embedding learning (Liu et al. 2019; Zhai et al. 2019), embedding for road segments (Deng et al. 2016; Liu et al. 2017), and embedding for spatial location distributions (Jean et al. 2019; Mai et al. 2020). Efforts in GIScience have also been made for developing spatially explicit models. For example, regionalization is a spatial constraint clustering procedure that groups areal objects into spatially contiguous homogeneous regions. Spatial constraints are explicitly considered in graph-based partition algorithms such as SKATER via minimum spanning tree and SKATER-CON via a set of random spanning trees (Assunção 2006; Aydin et al. 2018). For example, Figure 2 shows the spatial adjacency connectivity graph of counties in the state of Georgia, U.S. and illustrates the results of spatially constrained multivariate clustering analysis in ArcGIS 10. 7 using demographic and socioeconomic variables in Georgia counties.

Furthermore, spatial neighboring features can be added into feature space to improve the performance of predictive ML models and spatial statistical models (Anselin 1980; Fotheringham et al. 2003). Also, utilizing the rich information embedded in spatial contexts can improve the classification of place types from images of their facades and interiors and can outperform state-of-the-art computer-vision deep learning models (Yan et al. 2018). Depending on the consideration of first-order nonstationary effects (spatial heterogeneity) and second-order stationary effects (spatial autocorrelation), several PCA methods are classified into nonspatial and spatial-PCA approaches when applying PCA for dimension reduction and geospatial analysis (Demšar et al. 2013). Spatial smoothing kernels have been adopted to model the spatial nonstationarity in the relationship between target spatial distributed variables (e.g., PM2.5 concentrations) and predictor variables. A geographically-weighted gradient boosting machine (GW-GBM) was developed by improving traditional GBM through building spatial smoothing kernels to weigh the loss function and was applied for better spatiotemporal prediction of continuous daily PM2.5 concentrations across China (Zhan et al. 2017).

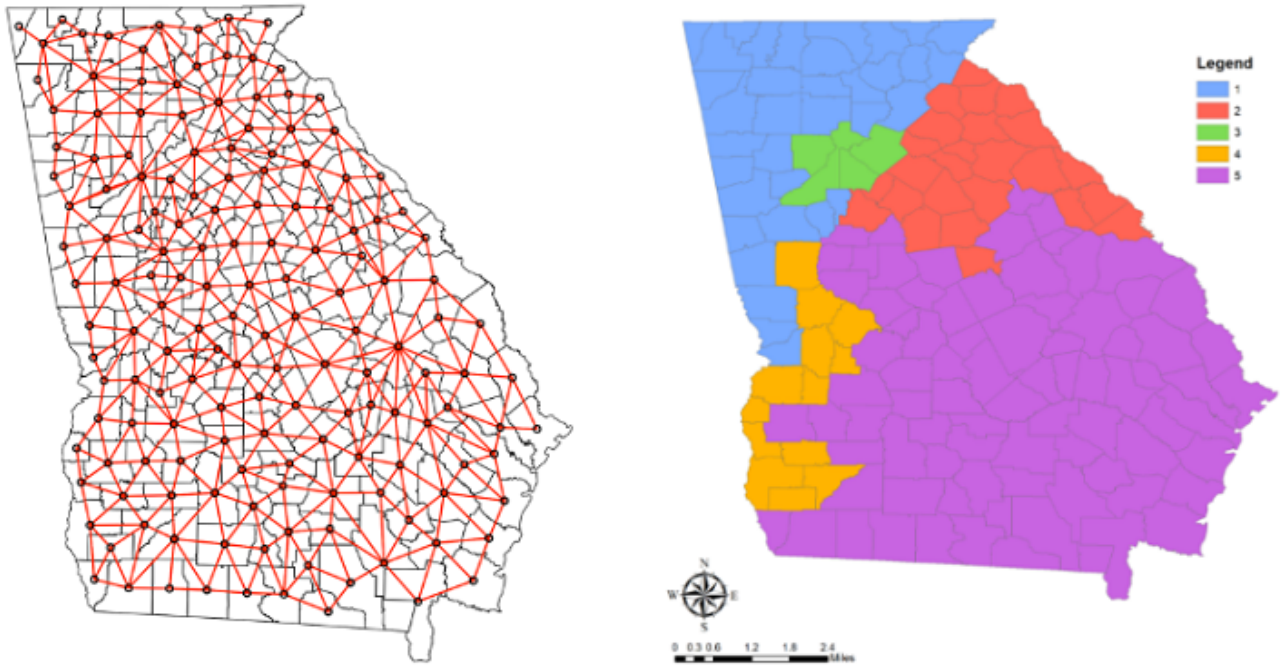


Figure 2: (left The spatial adjacency connectivity graph of counties in the state of Georgia, U.S.; (right) the results of spatially constrained multivariate clustering analysis using demographic and socioeconomic variables in Georgia counties. Maps source: authors.

3. ML Workflows and Applications

From a pragmatic perspective, machine learning is essentially an autonomous and self-learning workflow that has a feedback loop in which a learning process is used to train a ML model to find patterns and make predictions about new data. The steps of a ML workflow will vary according to the type of learning process being used to solve a classification, regression or a clustering problem. Figure 3 shows an overview of the main steps of a ML workflow for supervised/unsupervised learning. It starts with the study of the domain problem and then prepare the data accordingly. Oftentimes, the raw data require preprocessing for removing noise and performing data transformation, aggregation and contextualization. Then we partition the data into training, validation, and testing data sets. After that, the ML model is trained using input training data. The ML models are usually evaluated based on accuracy or error rate.

The ML model's hyperparameters are then set by the data scientist, and the other parameters can be further tuned and optimized in an iterative fashion. For example, a Deep Neural Network (DNN) consists of processing nodes (neurons), each with an operation performed on data as it travels through the multi-layer network (Goodfellow et al. 2016). Some hyperparameters such as the number of hidden layers and the number of nodes may be a-priori set differently across applications. When a DNN is trained, each node has a weight parameter that tells a model how much impact it has on the prediction results. When a model accuracy is acceptable for the classification problem, the trained ML model can be deployed in an inference server (e.g., Amazon SageMaker) or hardware (e.g., NVIDIA Jetson Nano Developer Kit) to digest online data and produce prediction results. A robust ML

model is expected to perform well on both training and testing data. It is worth noting that we may have nonrepresentative or insufficient quality of training data in practice and derive a poor ML model (Géron 2017). The model overfitting and underfitting issues require attentions when applying ML algorithms in practical applications (Mohri et al. 2018).



Figure 2. General workflow for supervised/unsupervised learning. Image source: authors. Image source: authors.

One of the pioneer studies on automated ML workflows in geography includes the Geographical Analysis Machine (GAM) for automated analysis of point data. Without prior information or particular location-specific hypotheses, the GAM was able to identify spatial clustering patterns of data on cancer in northern England (Openshaw, 1987). With the increasing popularity of data-driven approaches in geography, a variety of ML workflows have been applied in geospatial knowledge discovery and predictive modeling (Miller and Goodchild 2015). The ML workflows are preferred due to the challenges such as the data volume and the uncertainty arising with Big Geo-Data (Gao 2017; Yang et al. 2017). Big Geo-Data is used to describe the phenomenon that large volumes of georeferenced data (including structured, semi-structured, and unstructured data) about various aspects of the Earth environment, the society and human-land interactions are captured by millions of environmental and human sensors in a variety of formats such as remote sensing imageries, crowdsourced maps, social media (tweets, photos and videos), transportation smart card transactions, mobile phone data, location-based social networks, ride-sharing and GPS trajectories (Liu et al. 2015; Janowicz et al. 2019). Amount of studies have utilized these emerging big data sources with various ML models for dynamic population mapping, urban functions and land use inference (Vatsavai et al. 2005; Toole et al. 2012; Pei et al. 2014; Tu et al. 2017; Gao et al. 2017; Herfort et al. 2019; Li et al. 2019; Xu et al. 2019).

Recent advancement in machine learning and deep learning has revolutionized multiple domains in both scientific and practical ways (Jordan & Mitchell 2015; LeCun et al. 2015). Researchers argue that the integration of spatiotemporal features with deep learning models offers capabilities for better understanding of big data-driven and physical process-based Earth system science (Reichstein et al. 2019). Table 2 summarizes some examples of ML applications with key references in land use, soil mapping and environmental susceptibility, transportation, smart cities and social sensing, public health, crime analysis, surveillance and safety, just to name a few. In addition, some recent developed machine learning and particularly deep learning methods are also motivated by GIS, geography, cartography, and spatial statistics, such as deep compositional spatial models (Zammit-Mangion et al. 2019), spatially conditioned generative adversarial nets (Klemmer et al. 2019), Geo-GAN with reconstruction and style losses (Ganguli et al. 2019), vector representation learning of remote sensing data (Jean et al. 2019), and multi-scale location representation learning for spatial feature distributions (Mai et al. 2020).

Table 2. Applications Domains of Machine Learning

Domain	Used ML Algorithms	References
Land use, soil mapping and environmental susceptibility	KNN, multi-linear regression, fuzzy c-means clustering, hierarchical clustering, SVM, random forest, gradient boosting, multinomial logistic regression, expectation-maximization-based semi-supervised classification	Vatsavai et al. (2005), Marjanović et al. (2011), Toole et al. (2012), Pradhan (2013), Pei et al.(2014), Pham et al. (2016), Hengl et al. (2017), Tu et al. (2017), Chen et al. (2018), Taalab et al. (2018), Zhu et al. (2018), Du et al. (2019)
Transportation	KNN, SVM, random forest, Artificial Neural Networks (ANN), stacked-autoencoder neural networks (SAE), Long short-term memory neural networks (LSTM), graph convolutional network (GCN)	Vlahogianni et al. (2015), Lv et al. (2015), Ma et al. (2015), Liu et al. (2017), Polson and Sokolov (2017), Zhao et al. (2018), Cao and Wachowicz (2019); Gao et al. (2019), Maduako and Wachowicz (2019), Ren et al. (2019), Zhang et al. (2019)
Smart cities and social sensing	regularized linear regression, KNN, Bayesian models, SVM, random forest, topic models, gradient boosting trees, reinforcement learning, deep learning methods	Batty et al. (2012), McFarlane (2011), Liu et al. (2015), Jean et al. (2016), Regalia et al. (2016), Gao et al. (2017), Mohammadi & Al-Fuqaha (2018), Zhang et al. (2018), Dong et al. (2019), Herfort et al. (2019), Wu et al. (2019), Xu et al. (2019)
Public health	logistic regression, SVM, classification and regression trees (CART), random forest, DNN	Lee et al. (2010), Qu et al. (2011), Jensen et al. (2012), Allen et al. (2016), Gulshan et al. (2016), Ravì et al. (2017), Zhan et al. (2017), Mooney & Pejaver (2018)
Crime analysis, surveillance, and safety	DBSCAN, Spatiotemporal clustering, decision trees, SVM, ANN, DNN	Morris & Trivedi (2008), Hassani et al. (2016)

Finally, selecting the evaluation metrics for assessing ML models is an important step of any workflow. Many metrics have been proposed in different applications, and searching for a single metric may not be sufficient to give you insight on the accuracy of an ML model that has been used. Therefore, a subset of metrics are usually proposed to provide a tangible evaluation of ML models. The most common evaluation metrics used in the following ML models are one of the following (See Minaee (2019) for more information):

- Classification Metrics (accuracy, precision, recall, F1-score, ROC, AUC)
- Regression Metrics (MSE, MAE)
- Ranking Metrics (MRR, DCG, NDCG)
- Deep Learning Related Metrics (Inception Score, Frechet Inception Distance)

4. Challenges and a Vision for the Future

The challenges in learning from data have led to a revolution in Geospatial Data Science. The ongoing explosion and increasing availability of geospatial big data is driving the development of more specialized and data specific ML models. However, current ML Models still take considerable amount of time and computational resources for being trained and yet failing to predict at all. Automated ML workflows or even well accepted rules of thumb for a learning process are still out of reach. Critics about the adoption of machine learning arise with a key warning that ML models may contribute to a growing replicability and

reproducibility crisis in science (National Academies of Sciences, Engineering, and Medicine, 2019). Future developments should assess the uncertainty and reproducibility of their predictions using machine learning. Interpreting “black box” ML prediction models also requires the understanding of a domain knowledge. Automated discovery and explanatory models may be developed from geospatial data using inductive process modeling based on fragments of domain understanding (Gahegan 2020). The location uncertainty and the spatial heterogeneity along with other characteristics in geospatial data make this issue more prominent when researchers apply machine learning techniques in Geospatial Data Science. Future studies and developments are needed to address such concerns. Moreover, applying ML to spatial data is not a one-way street. Location is a key to integrate and synthesize multi-source data layers; geographic domain knowledge and spatial concepts can help contribute to develop different contextual spaces (i.e. mobility space, social space, and event space) which will play an important role in the development of ML models in general.

Algorithm transparency has emerged as one research challenge due to strong concerns about the inherited high risk of bias in data and algorithms being used in automated ML workflows. It will require broader steps to resolve them, but current research is pointing out to the need to understand how feature spaces and social places that are used to collect training data can influence the behavior of the predictive models. In the United States, there are already initiatives such as the AI Now Institute in which New York University and the Algorithmic Justice League with the help of MIT Media Lab that are warning about the power of biased algorithms and their socio-spatial implications.

The challenge of labeling training data remains the main source of error in machine learning since it will continue to be manually carried out by humans for many years to come. Recently leading researchers have stressed the need for a new theory for machine learning, pointing out that the human brain learns without the need of all that labeled data to reach a conclusion. Geoffrey Hinton goes far as to say, “My view is throw it all away and start again.” The time is ripe for a new theory in Geospatial Data Science.

References

Allen, C., Tsou, M. H., Aslam, A., Nagel, A., & Gawron, J. M. (2016). Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. PloS one, 11(7).

Anselin, L. (1980) Estimation Methods for Spatial Autoregressive Structures, Regional Science Dissertation and Monograph Series No. 8, Cornell University, Ithaca, NY.

Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20(7), 797-811.

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A Survey. Computer Networks, 54(15), 2787-2805.

Aydin, O., Janikas, M. V., Assunção, R., & Lee, T. H. (2018, November). SKATER-CON:

Unsupervised Regionalization via Stochastic Tree Partitioning within a Consensus Framework Using Random Spanning Trees. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (pp. 33-42). ACM.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., ... & Portugali, Y. (2012). Smart cities of the future. The European Physical Journal Special Topics, 214(1), 481-518.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A Review and New Perspectives. IEEE Transactions and Machine Intelligence, 35(8), 1798-1828.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer, New York.

Campello, R.J.G.B., Moulavi, D., Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science(), vol 7819. Springer, Berlin, Heidelberg

Cao, H. and Wachowicz, M. (2019). An edge-fog-cloud architecture of streaming analytics for Internet of Things applications. Sensors, Special Issue on Issue Edge/Fog/Cloud Computing in the Internet of Things; Velasco, L and Ruiz, M. (Eds), 19:3594

Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., Zhu, A., Pei, X., & Duan, Z. (2018). Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. Science of the Total Environment, 626, 1121-1135.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal Component Analysis on Spatial Data: An Overview. Annals of the Association of American Geographers, 103(1), 106-128.

Deng, D., Shahabi, C., Demiryurek, U., Zhu, L., Yu, R., & Liu, Y. (2016, August). Latent Space Model for Road Networks to Predict Time-Varying Traffic. In Proceedings of the 22nd KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1525-1534).

Dong, L., Ratti, C., & Zheng, S. (2019). Predicting neighborhoods' socioeconomic attributes using restaurant data. Proceedings of the National Academy of Sciences, 116(31), 15447-15452.

Du, F., Zhu, A. X., Liu, J., & Yang, L. (2019). Predictive mapping with small field sample data using semi-supervised machine learning. Transactions in GIS, 24(2):315-331.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD'96, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (Vol. 96, No. 34, pp. 226-231).

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The Journal of Machine Learning Research, 15(1), 3133-3181.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons.

Gahegan, M. (2020). Fourth paradigm GIScience? Prospects for automated discovery and explanation from data. International Journal of Geographical Information Science, 34(1), 1-21.

Ganguli, S., Garzon, P., & Glaser, N. (2019). GeoGAN: A conditional GAN with reconstruction and style loss to generate standard layer of maps from satellite images. arXiv.

Gao, S. (2017). Big Geo-Data. In Laurie A. Schintler and Connie L. McNeely (Eds): Encyclopedia of Big Data, Springer.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. Transactions in GIS, 21(3), 446-467.

Gao, S., Li, M., Liang, Y., Marks, J., Kang, Y., & Li, M. (2019). Predicting the spatiotemporal legality of on-street parking using open data and machine learning. Annals of GIS, 25(4), 299-312.

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.

Goodfellow I., Bengio Y., & Courville A. (2016). Deep Learning. Cambridge, MA: The MIT Press.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Kim, R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Journal of the American Medical Association (JAMA) 316(22), 2402-2410.

Hand, D. J. (2016). Measurement: A Very Short Introduction. Oxford University Press.

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(3), 139-154.

Hastie, T, Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. New York: Springer Series in Statistics.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ... & Guevara, M. A. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS one, 12(2), e0169748.

Herfort, B., Li, H., Fendrich, S., Lautenbach, S., & Zipf, A. (2019). Mapping Human Settlements with Higher Accuracy and Less Volunteer Efforts by Combining Crowdsourcing and Deep Learning. Remote Sensing, 11(15), 1799.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y., & Bhaduri, B. (2020). GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. International Journal of Geographical Information Science, 0(0), 1-13.

Janowicz, K., McKenzie, G., Hu, Y., Zhu, R., & Gao, S. (2019). Using Semantic Signatures for Social Sensing in Urban Environments. In Mobility Patterns, Big Data and Transport Analytics (pp. 31-54). Elsevier.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., & Ermon, S. (2019, July). Tile2Vec: Unsupervised representation learning for spatially distributed data. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 3967-3974).

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.

Klemmer, K., Koshiyama, A., & Flennerhag, S. (2019). Augmenting correlation structures in spatial data using deep generative models. arXiv preprint.

LeCun, Y. L., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in Medicine, 29(3), 337-346.

Li, M., Gao, S., Lu, F., & Zhang, H. (2019). Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data. Computers, Environment and Urban Systems, 77, 101346, 1-10.

Liu, K., Gao, S., Qiu, P., Liu, X., Yan, B., & Lu, F. (2017). Road2Vec: Measuring Traffic Interactions in Urban Road System from Massive Travel Routes. ISPRS International Journal of Geo-Information, 6(11), 321.

Liu, X., Andris, C., & Rahimi, S. (2019). Place niche and its regional variability: Measuring

spatial context patterns for points of interest with representation learning. Computers, Environment and Urban Systems, 75, 146-160.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., ... Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. Annals of the Association of American Geographers, 105(3), 512–530.

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2015). Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems, 16(2), 865-873.

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies, 54, 187-197.

Maduako, I. and Wachowicz, M. (2019). A Space-Time Varying Graph for Modelling Places and Events in a Network. International Journal of Geographical Information Science. 33(10): 1915-1935.

Mai, G., Janowicz, K., Hu, Y., & Gao, S. (2018). ADCN: An anisotropic density-based clustering algorithm for discovering spatial point patterns with noise. Transactions in GIS, 22(1), 348-369.

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., Lao, N. (2020) Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells. The Eighth International Conference on Learning Representations (ICLR 2020). 1-13.

Marjanović, M., Kovačević, M., Bajat, B., & Vozenílek, V. (2011). Landslide susceptibility assessment using SVM machine learning algorithm. Engineering Geology, 123(3), 225-234.

McFarlane, C. (2011). The city as a machine for learning. Transactions of the Institute of British Geographers, 36(3), 360-376.

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. ACM Computing Surveys, 45(1), 10.

Michie, D. (1968). “Memo” Functions and Machine Learning. Nature, 218(5136), 19.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K. R. (1999, August). Fisher discriminant analysis with kernels. In Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468), Madison, WI, USA. pp. 41-48.

Miller, H. J. and Goodchild, M. F. (2016) Data-Driven Geography. GeoJournal, 80: 449-461.

Minaee, S. (2019). 20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics. Towards Data Science.

Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), 1979-1993.

Mohammadi, M., & Al-Fuqaha, A. (2018). Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges. IEEE Communications Magazine, 56(2), 94-101.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning, 2nd Edition. Cambridge, MA: The MIT Press.

Mooney, S. J., & Pejaver, V. (2018). Big Data in Public Health: Terminology, Machine Learning, and Privacy. Annual Review of Public Health, 39, 95-112.

Morris, B. T., & Trivedi, M. M. (2008). A Survey of Vision-Based Trajectory Learning and Analysis for Surveillance. IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 8, pp. 1114-1127.

NASEM, National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press.

Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. International Journal of Geographical Information System, 1(4), 335-358.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science, 28(9), 1988-2007.

Pham, B. T., Pradhan, B., Bui, D. T., Prakash, I., & Dholakia, M. B. (2016). A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environmental Modelling & Software, 84, 240-250.

Pilevar, A. H., & Sukumar, M. (2005). GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. Pattern Recognition Letters, 26(7), 999-1010.

Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. Transportation Research Part C: Emerging Technologies, 79, 1-17.

Qu, H. Q., Li, Q., Rentfro, A. R., Fisher-Hoch, S. P., & McCormick, J. B. (2011). The definition of insulin resistance using HOMA-IR for Americans of Mexican descent using machine learning. PloS one, 6(6), e21041.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z.

(2017). Deep Learning for Health Informatics. IEEE journal of biomedical and health informatics, 21(1), 4-21.

Regalia, B., McKenzie, G., Gao, S., & Janowicz, K. (2016). Crowdsensing smart ambient environments and services. Transactions in GIS, 20(3), 382-398.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195-204.

Ren, Y., Cheng, T., & Zhang, Y. (2019). Deep spatio-temporal residual neural networks for road-network-based data modeling. International Journal of Geographical Information Science, 33(9), 1894-1912.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. II—Recent progress. IBM Journal of research and development, 11(6), 601-617.

Singleton, A. and Arribas-Bel, D. (2019). Geographic Data Science. Geographical Analysis, 53(1): 61-75.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. Science 103, 677-680.

Taalab, K., Cheng, T., & Zhang, Y. (2018). Mapping landslide susceptibility and types using Random Forest. Big Earth Data, 2(2), 159-178.

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012, August). Inferring land use from mobile phone activity. UrbComp '12: Proceedings of the ACM SIGKDD International Workshop on Urban Computing (pp. 1-8).

Trillos, N. G., & Murray, R. (2017). A new analytical approach to consistency and overfitting in regularized empirical risk minimization. European Journal of Applied Mathematics, 28(6), 886-921.

Tu, W., Cao, J., Yue, Y., Shaw, S. L., Zhou, M., Wang, Z., ... & Li, Q. (2017). Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. International Journal of Geographical Information Science, 31(12), 2331-2358.

Vatsavai, R. R., Shekhar, S., & Burk, T. E. (2005, November). A semi-supervised learning method for remote sensing data mining. In 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05) (pp. 5-pp). Hong Kong, China.

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. Transportation Research Part C: Emerging Technologies, 43, 3-19.

Volpi, M., & Tuia, D. (2018). Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 144, 48-60.

Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X., & Liu, Y. (2019). Inferring demographics from human trajectories and geographical context. Computers, Environment and Urban Systems, 77, 101368, 1-11.

Xu, Y., Chen, D., Zhang, X., Tu, W., Chen, Y., Shen, Y., & Ratti, C. (2019). Unravel the landscape and pulses of cycling activities from a dockless bike-sharing system. Computers, Environment and Urban Systems, 75, 184-203.

Yan, B., Janowicz, K., Mai, G., & Gao, S. (2017, November). From ITDL to Place2Vec: Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts. SIGSPATIAL '17: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pages 1-10.

Yan, B., Janowicz, K., Mai, G., & Zhu, R. (2018). xnet+ sc: Classifying places based on images by incorporating spatial contexts. In 10th International Conference on Geographic Information Science (GIScience 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 17, 1–15.

Yan, B., Janowicz, K., Mai, G., & Zhu, R. (2019). A spatially explicit reinforcement learning model for geographic knowledge graph summarization. Transactions in GIS, 23(3), 620-640.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. International Journal of Digital Earth, 10(1), 13-53.

Zammit-Mangion, A., Ng, T. L. J., Vu, Q., & Filippone, M. (2022). Deep Compositional Spatial Models. Journal of the American Statistical Association 117(540).

Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z. R., & Gu, C. (2019). Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. Computers, Environment and Urban Systems, 74, 1-12.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., ... & Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM2. 5 concentrations across China using a spatially explicit machine learning algorithm. Atmospheric Environment, 155, 129-139.

Zhang, C., Zhang, K., Yuan, Q., Peng, H., Zheng, Y., Hanratty, T., ... & Han, J. (2017, April). Regions, Periods, Activities: Uncovering Urban Dynamics via Cross-Modal Representation Learning. In Proceedings of the 26th International Conference on World Wide Web (pp. 361-370).

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning, 180, 148-160.

Zhang, Y., Cheng, T., Ren, Y., & Xie, K. (2019). A novel residual graph convolution deep learning model for short-term network-based traffic forecasting. International Journal of Geographical Information Science, 34(5) 1-27.

Zhao, L., Song, Y., Zhang, C., Liu, Y., Wang, P., Lin, T., ... & Li, H. (2019). T-GCN: A temporal graph convolutional network for traffic prediction. IEEE Transactions on Intelligent Transportation Systems, 21(9):1-11.

Zhu, A. X., Lu, G., Liu, J., Qin, C. J., and Zhou, C. (2018). Spatial prediction based on the Third Law of Geography. Annals of GIS, 24(4):225–240.