ABSTRACT

Challenges in KDD and ML for Sustainable Development

LAURE BERTI-EQUILLE, Space Observation, Models & Actionable Science, Montpellier, Occitanie, France

DAVID DAO, Swiss Federal Institute of Technology, Zurich, Zurich, ZH, Switzerland

STEFANO ERMON, Stanford University, Stanford, CA, United States

BEDHARTA GOSWAMI, University of Tübingen, Tubingen, Baden-Wurttemberg, Germany

Published: 14 August 2021

KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 14 - 18, 2021
Virtual Event, Singapore

Conference Sponsors:
SIGMOD
SIGKDD

KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (August 2021)
https://doi.org/10.1145/3447548.3470798
ISBN: 9781450383325

Challenges in KDD and ML for Sustainable Development

Laure Berti-Equille
IRD ESPACE DEV
Montpellier, France
laure.berti@ird.fr

David Dao
ETH Zürich
Zürich, Switzerland
david.dao@inf.ethz.ch

Stefano Ermon
Stanford University
Stanford, CA, USA
ermon@cs.stanford.edu

Bedharta Goswami
University of Tübingen
Tübingen, Germany
bedartha.goswami@uni-tuebingen.de

Abstract

Artificial Intelligence and machine learning techniques can offer powerful tools for addressing the greatest challenges facing humanity and helping society adapt to a rapidly changing climate, respond to disasters and pandemic crisis, and reach the United Nations (UN) Sustainable Development Goals (SDGs) by 2030. In recent approaches for mitigation and adaptation, data analytics and ML are only one part of the solution that requires interdisciplinary and methodological research and innovations. For example, challenges include multi-modal and multi-source data fusion to combine satellite imagery with other relevant data, handling noisy and missing ground data at various spatio-temporal scales, and ensembling multiple physical and ML models to improve prediction accuracy. Despite recognized successes, there are many areas where ML is not applicable, performs poorly or gives insights that are not actionable. This tutorial will survey the recent and significant contributions in KDD and ML for sustainable development and will highlight current challenges that need to be addressed to transform and equip engaged sustainability science with robust ML-based tools to support actionable decision-making for a more sustainable future.

CCS Concepts

• **Social and professional topics** → **Sustainability**; • **Computing methodologies** → **Machine learning approaches**; • **Applied computing** → **Earth and atmospheric sciences**; **Environmental sciences**; *Decision analysis*.

ACM Reference Format:
Laure Berti-Equille, David Dao, Stefano Ermon, and Bedharta Goswami. 2021. Challenges in KDD and ML for Sustainable Development. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3470798

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3470798

1 Context and Motivation

The United Nations' Sustainable Development Goals (SDGs) are a set of 17 goals adopted by UN members states in 2015 to help create a safer, more sustainable, and prosperous planet. The SDGs, which compass 169 individual targets, form part of the UN's 2030 Agenda for Sustainable Development and are meant to "stimulate action over the next fifteen years in areas of critical importance for humanity and the planet". In-line with these salutary goals, the tutorial will attempt to examine how ML and data mining have contributed so far and advanced the state-of-the-art as applied sciences. This tutorial will not cover exhaustively all the SDGs but rather present successes and limitations of ML applied to a selected set of use cases under the umbrella of three SGDs: (1) exploiting Earth Observation data and satellite imagery to estimate poverty [5, 8, 10] related to SDG #1 No Poverty; (2) ML-based climate data analytics [6] related to SDG #13 Climate Action; and (3) ML-based monitoring for forest and biodiversity conservation [4] related to SDG #15 Life on Land.
Resources. The slide deck and videos are available at:
https://laureberti.github.io/KDD2021_Tutorial/.

2 Tutorial Outline

The tutorial will start with an introductory overview of the relevant concepts and methods in Data Analytics and Machine Learning applied to Sustainable Development (SD) with a SWOT analysis. We will explore the use of data science and ML techniques as tools to integrate multi-modal, multi-source data and human multidisciplinary expertise. We will reformulate a set of SD-related questions into formal ML problem statements and present some illustrative examples and real-world study cases from various application domains related to climate action, clean and sustainable energy, and biodiversity conservation [1]. We will provide an overview of the opportunities and limitations, alongside with computational, technical, and operational challenges associated with ML applied to sustainability development. Next, we will present the main challenges of ML applied to SD by articulating the presentation on the following generic pipeline: (1) Understand the input and validation data, actors, and the target SD goal; (2) Collect, integrate, and prepare multi-source and multi-modal data sets; (3) Select features, ML models/architectures, and parameters; (4) Include multidisciplinary expertise with Human-In-the-Loop (HIL) and user interaction; and

(5) Validate and evaluate the actionability, transferability, and reproductibility of the pipeline to other SD settings. Next, we deep dive into three SD applications that required the adaptation and design of new ML methods to address various sets of technical and theoretical challenges.

ML and Satellite Imagery to Estimate Poverty. Recent technological developments are creating new data streams that contain a wealth of information relevant to sustainable development goals [13]. Modern AI techniques have the potential to yield accurate, inexpensive, and highly scalable models to inform research and policy. A key challenge, however, is the lack of large quantities of labeled data that often characterize successful machine learning applications. We present new approaches for learning useful spatio-temporal models in contexts where labeled training data is scarce or not available at all [2, 9, 15]. We show applications to predict and map poverty in developing countries, monitor agricultural productivity and food security outcomes, and map infrastructure access in Africa. The proposed methods can reliably predict economic well-being using only high-resolution satellite imagery. Because images are passively collected in every corner of the world, the methods can provide timely and accurate measurements in a very scalable end economic way, and could revolutionize efforts towards global poverty eradication.

ML-based Climate Data Analytics. Next, we present an overview of different machine learning based approaches that have been used in climate data analysis. First, we look at classical approaches such as principal component analysis (PCA) along with its nonlinear extensions, which include kernel based methods as well as autoencoder based approaches. We also discuss correlation-based hierarchical clustering as an alternative to PCA for identifying a lower dimensional representation of spatio-temporal climate data sets. Then, we introduce climate networks and a sparse representation of functional relations between spatially distributed climate time series, and we look at how climate networks have been used to detect, quantify, and predict complex climate phenomena. In a second part, we introduce the fundamental paradigm of paleoclimate proxy measurements and the challenges that arise due to dating uncertainties. We present a Bayesian estimation approach of paleo-proxy uncertainties and its numerical approximation. This allow us to formulate a new representation of time series, as a sequence of probability density functions (PDFs) in lieu of point-like measurements. Finally, we use the *time series as PDF sequence* representation to show how recurrence plots can be used to detect abrupt transitions in time series with uncertainties.

ML to Help Restore the Natural World. Land use and its evolution play a critical role in our climate [7], taking up about a quarter of annual anthropogenic emissions of greenhouse gases (GHGs) during 2007-2016 [12]. In addition to being a key driver of global warming, careless land use is also destroying valuable ecosystem services and is threatening the livelihood for local populations and a multitude of species. Major conservation and restoration efforts are underway to mitigate and safeguard against these losses, and to highlight the urgency of the issue, 2021-2030 has been declared the "UN Decade on Ecosystem Restoration". However, we cannot preserve what we cannot measure. ML plays a significant role in responding to this critical call for action and can accelerate the conservation and sustainable use of our natural world. We first present the background on the importance of the natural world on climate change and the current biodiversity crisis. Next, we will give an overview of current MRV (Monitoring, Reporting, and Verification) pipelines and present a case study of how AI and ML can fit into and scale the existing MRV pipelines [3, 11, 14].

Significant efforts must still be spent to adapt traditional KDD and ML techniques to solve environmental and climate-related problems. We need solutions and pathways leading to robust mitigation of dangerous anthropogenic climate change. Data science and ML models can help in identifying such pathways toward a sustainable future and can be used for informing the policymakers and the wider public. Leveraging ML for SD is a vast, challenging, and still understudied area for which the KDD community has a role to play.

References

[1] F. Amato, F. Guignard, and e. a. S. Robert. 2020. A novel framework for spatio-temporal prediction of environmental data using deep learning. *Sci. Rep.* 22243, 10 (2020). https://www.nature.com/articles/s41598-020-79148-7

[2] K. Ayush, B. Uzkent, M. Burke, D. Lobell, and S. Ermon. 2020. Generating Interpretable Poverty Maps using Object Detection in Satellite Images. In *Proceedings of IJCAI 2020*. 4410–4416.

[3] D. Dao, C. Cang, C. Fung, M. Zhang, N. Pawlowski, R. Gonzales, N. Beglinger, and C. Zhang. 2019. GainForest: Scaling Climate Finance for Forest Conservation using Interpretable Machine Learning on Satellite Imagery. In *Proceedings of the ICML Climate Change AI workshop 2019*.

[4] D. Dao and J. Rausch. 2019. GeoLabels: Towards Efficient Ecosystem Monitoring using Data Programming on Geospatial Information. In *NeurIPS Climate Change workshop 2019*.

[5] C. D. Elvidge, P. C. Sutton, T. Ghosh, B. T. Tuttle, K. E. Baugh, B. Bhaduri, and E. Bright. 2009. A global poverty map derived from satellite data. *Computers & Geosciences* 8 (Aug. 2009), 1652–1660.

[6] B. Goswami, N. Boers, A. Rheinwalt, N. Marwan, J. Heitzig, S. F. M. Breitenbach, and J. Kurths. 2018. Abrupt transitions in time series with uncertainties. *Nature Communications* 9, 1 (2018), 48. https://doi.org/10.1038/s41467-017-02456-6

[7] P. Helber, B. Bischke, A. Dengel, and D. Borth. 2018. Introducing EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. In *Proceedings of IGARSS 2018*. IEEE, 204–207.

[8] R. Jarry, M. Chaumont, L. Berti-Équille, and G. Subsol. 2021. Assessment of CNN-based Methods for Poverty Estimation from Satellite Images. In *Proceedings of the 11th IAPR International Workshop on Pattern Recognition in Remote Sensing (PRRS) in conjunction with the International Conference on Pattern Recognition (ICPR 2020)*. Milan, Italy.

[9] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (Aug. 2016).

[10] L. Kondmann and X. X. Zhu. 2020. Measuring Changes in Poverty with Deep Learning and Satellite Images. *Proceedings of ICLR 2020* (2020).

[11] S. Santamaria, D. Dao, B. Lütjens, and C. Zhang. 2020. TrueBranch: Metric Learning-based Verification of Forest Conservation Projects. arXiv:2004.09725 [cs.CV]

[12] P. Shukla, J. Skea, E. C. Buendia, V. Masson-Delmotte, H.-O. Pörtne, D. Robert, P. Zhaian, R. Slade, S. Connors, R. van Diemen, M. Ferrat, E. Haughey, S. Luz, S. Neogi, M. Pathak, J. Petzold, J. P. Pereira, P. Vyas, E. Huntley, K. Kissick, M. Belkacemi, and J. Malley (Eds.). 2019. IPCC. 2019: Summary for policymakers. *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems* (2019), 7–11.

[13] Thematic Research Network on Data and Statistics. 2020. *Leaving no one off the map: A guide for gridded population data for sustainable development, Thematic Research Network on Data and Statistics*. Technical Report. UN Sustainable Development Solutions Network.

[14] B. G. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White. 2019. Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks. *Remote Sensing* 11, 11 (2019), 1309.

[15] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* (May 2020).