Wildfire Smoke and Air Quality: How Machine Learning Can Guide Forest Management

Lorenzo Tomaselli [1] **Coty Jen** [2] **Ann B. Lee** [1]

[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]Department of Chemical Engineering, Carnegie Mellon University
ltomasel@andrew.cmu.edu

Abstract

Prescribed burns are currently the most effective method of reducing the risk of widespread wildfires, but a largely missing component in forest management is knowing which fuels one can safely burn to minimize exposure to toxic smoke. Here we show how machine learning, such as spectral clustering and manifold learning, can provide interpretable representations and powerful tools for differentiating between smoke types, hence providing forest managers with vital information on effective strategies to reduce climate-induced wildfires while minimizing production of harmful smoke.

1 Introduction

As evident from the past decade, wildfires in the western US are becoming larger, deadlier, and more frequent due to worsening climate conditions, causing phenomenons such as recurring droughts, heat waves, and earlier springs [3; 11; 19]. In an effort to reduce wildfire risk, state and local governments are pushing to manage forests via regular prescribed burning, i.e., controlled, low-temperature burns to reduce the amount of fuel in a forest. Although prescribed burning conducted every few years will reduce wildfire risk due to climate change, populations near heavily forested areas will experience frequent smoke episodes. Previous research has observed that smoke produced from different burn conditions or from burning different fuels, such as whole trees and man-made structures in wildfires versus ground cover duff in prescribed burns, is chemically distinct [1; 9]. Specifically, we have previously demonstrated that burning a specific shrub, manzanita, can release enormous amounts of a toxic compound known as hydroquinone into the air [10]. Other smoke compounds that describe a specific fuel or burn condition may exist but there is currently no technique that can extract this information from the complex chemical composition. Consequently, current air quality models treat all smoke as composed of about ten common compounds (i.e., produced from burning any vegetation) and overlooks differences between thousands of other compounds [20]. The lack of more fuel-specific smoke information prevents forest managers from developing an effective strategy to reduce climate-induced wildfire risk while minimizing air quality damage.

A first step toward understanding the impact of wildland burns, including wildfires and prescribed burns, on air quality is through *chemical fingerprinting* — laboratory analysis to quantify compounds in collected smoke samples from wildland burns. The output from advanced analytic platforms, such as two-dimensional gas chromatography (GC) and mass spectrometry (MS), is multi-dimensional and contains thousands of chemical compounds [5]. Comparing smoke samples and relating fingerprint data to their impacts on air quality, and subsequent exposure toxicity, is also an inherently difficult task because of complex synergism between numerous compounds and the variability in burn conditions. There is an urgent need for machine learning techniques that can differentiate between different types of smoke with interpretable and explainable results. We do not want to limit ourselves to

Tackling Climate Change with Machine Learning workshop at NeurIPS 2020.

only answering whether two smoke samples are different, but we aim at gaining insight as to how differences (if observed and statistically significant) relate to groups of compounds. In this way we can link smoke to fuel, burn conditions, and air quality impacts.

In this work, we describe how spectral clustering and manifold learning can provide interpretable representations and powerful tools for differentiating between smoke. The statistical challenge is that the data (such as mass spectra representing chemical compounds) are not only high-dimensional, but that each smoke sample is a weighted set of high-dimensional data. Furthermore, not all mass spectra are equally likely to occur; the underlying physical constraints naturally imply that the spectra, if seen as points in a higher-dimensional space, inherently possess "sparse" low-dimensional structure (clusters, manifolds, etc). The main question is how to design efficient representations and machine learning (ML) algorithms for smoke samples that take sparse structure into account, and that allow us to easily relate ML prediction results back to groups of chemical compounds. Here we present a new geometry-based metric between smokes as distributions over chemical compounds, which can serve as input to kernel ML and visualization algorithms.

2 Data

Our research team collected a total of 54 prescribed burn smoke samples at ground level and above the forest canopy ($\sim$100 m in the air). Burns were conducted at Blodgett Forest Research Station (BFRS, Georgetown, CA) in 2017. Additional $\sim$100 smoke samples will be collected during prescribed burns at BFRS in Nov. 2020. These field samples represent unknown fuel composite samples as we only know the approximate fuel composition prior to the prescribed burn. We will also conduct laboratory burns of fuels harvested from BFRS in order to separate chemical fingerprints of smoke produced from single and multicomponent fuel mixtures.

Smoke samples are analyzed using two-dimensional GC-MS with online derivatization [9]. This instrument separates compounds within smoke particles by first separating molecules by their boiling point and then by their polarity before detection via electron ionization with a time-of-flight mass spectrometer. A compound is described by its retention time for boiling point (i.e., the time needed to travel through the first GC column), retention time for polarity, and mass spectrum. Approximately 1000 unique compounds were separated and quantified in each smoke sample. We define the *chemical fingerprint* of a smoke sample as a weighted set of compounds $S = \{(\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_m, w_{\mathbf{x}_m})\}$, where $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^p$ represent $m \sim 1000$ chemical compounds in the sample as measured by MS, GC, or both, with dimension $p \sim 500$; the weights $w_{\mathbf{x}_1}, \ldots, w_{\mathbf{x}_m} \in [0, 1]$ with $\sum_{i=1}^m w_{\mathbf{x}_i} = 1$ represent emission intensities or mass concentrations (mg/m^3).

3 Interpretable Insights on Smoke via Geometry-Based Metrics

Aim 1: Building interpretable metrics that differentiate between smoke via manifold learning

Although the data $\mathbf{x} \in \mathbb{R}^p$ (for, e.g., mass spectra of chemical compounds) are high-dimensional, they have low intrinsic dimensionality due to physical constraints in the system. *How do we build interpretable metrics that reflect this structure with scalable ML algorithms for computing distances between smoke samples?* The toy example in Figure 1 illustrates our proposed approach for comparing

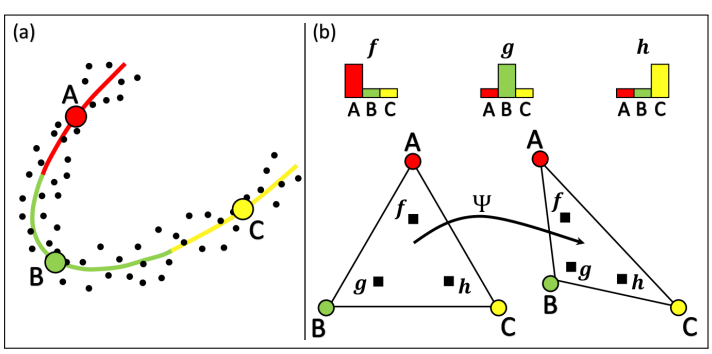

Figure 1: Toy example to illustrate Aim 1. In (a), we define a codebook for chemical compounds via spectral clustering. Panel (b), bottom left, illustrates that a distance metric that does not utilize data geometry will place histogram representations f, g, and h of smoke samples at equal distance, whereas the arrangement to the bottom right (implied by GDD) better reflects that f is closer to g than to h.

and analyzing sets of such high-dimensional data. Panel (a) shows data lying near a low-dimensional manifold. Via spectral clustering, using the diffusion K-means algorithm [13], we quantize the compound space according to geometric structure to create a common "codebook" with centroids A, B, and C. The natural geometry of the data suggests that points near A are closer to points near B than those around C. Panel (b), top, shows three histograms, f, g and h, for distributions over these centroids. The bottom left triangle illustrates how any metric that does not take data geometry into account views the location of the three histograms on a two-dimensional simplex; each of the bins corresponds to a vertex on an equilateral triangle. The arrangement of the bins and the histograms in the bottom right triangle better reflects the interbin relationships.

Our proposed method takes geometry into account by placing the bins (code words) in a Euclidean space defined by a diffusion map. The diffusion map is a rescaled Laplacian eigenmap [2], $\mathbf{x} \in \mathbb{R}^p \mapsto \Psi(\mathbf{x}) = (\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots, \psi_d(\mathbf{x})) \in \mathbb{R}^d$, where $\psi_1, \ldots, \psi_d$ are the first d eigenvectors (ordered by their eigenvalues) of a positive semi-definite "kernel"; in our settings, this kernel could for example be a (renormalized) Gaussian kernel based on cosine distances [18] between chemical compounds. In the diffusion framework, the rescaling is defined such that the Euclidean distance $\|\Psi(\mathbf{x}_i) - \Psi(\mathbf{x}_j)\|$ between any two compounds i and j in the map reflects the intrinsic connectivity structure of the data, defined by a Markov random walk (MRW) over the underlying (low-dimensional) data structure. The details of the construction are given in [13]; similar ideas can also be found in [4; 14] . The value of having an explicit diffusion metric is that it offers a principled way of linking spectral clustering and manifold learning to data compression and visualization.

Contribution. In this work, we extend the diffusion metric between high-dimensional data (chemical compounds) to a generalized diffusion distance (GDD) between weighted sets of such data (smoke samples). Let $\mathbf{f} = \{f_1, \ldots, f_K\}$ and $\mathbf{g} = \{g_1, \ldots, g_K\}$ be the histogram representations of any two fingerprints, say S and $\widetilde{S}$, where the bin locations are defined by the diffusion K-means centers $\{c_i\}_{i=1}^K$ with $K \sim 100$. One can show that the GDD between histograms can be computed very efficiently as $GDD(S, \widetilde{S}) = \left\| \sum_{i=1}^K (f_i - g_i) c_i \right\|$,[1] thus with linear complexity $O(K)$. In comparison, another common set metric such as the Earth Mover Distance (EMD)[16] has a typical algorithmic complexity of $O(K^3 \log K)$ and also does not take the intrinsic geometry of the data distribution into account. Finally, one can show (in prep) that in the limit of $K \to m$ (or no binning), a version of the GDD converges to the well-known Maximum Mean Discrepancy (MMD) [6]. The latter result can be used to determine the amount of compression (or the smallest K) that would give *maximum interpretability, computational efficiency* and *low variance* without sacrificing significant accuracy (or bias) relative to MMD.

Aim 2: Differentiating between smoke via kernel machine learning and statistical tests

GDD can serve as input to kernel-based machine learning algorithms [7; 17] for regression and classification of different types of smoke. Our BFRS data include smoke samples collected at air versus ground level. Smoke from the first category escapes the tree line and is more likely to affect communities living near a prescribed burn. Our proposed framework allows us to distinguish between the two different types of smoke, as well as quantify the composition of the smoke most likely to affect regional air quality. In addition to probabilistic classification, we propose a principled method of identifying whether differences in smoke samples are statistically significant. For the latter task, we will either employ the local regression test [12] based on a kernel logistic regression [21] with GDD, or alternatively, test for differences within each of the K partitions via a binomial test [8; 15].

Aim 3: Mapping back to compound space to visualize and explain results

Figure 2 illustrates our methodology on our initial BFRS samples. Once we have defined a common codebook for chemical compounds and represented each chemical fingerprint of smoke as a histogram over the code words (orange markers), we can visualize the smoke samples in diffusion space as points in the convex hull of the K "words". This inverse map back to compound space provides

[1] The original diffusion distances $D(\mathbf{x}_i, \mathbf{x}_j)$ between compounds i and j correspond to the transition probabilities of a MRW on a graph with the compounds as nodes and the edge weights given by a suitable similarity matrix; the above metric between individual compounds induces a new metric $GDD(S, \widetilde{S})$ between weighted sets $S = \{(\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_m, w_{\mathbf{x}_m})\}$ and $\widetilde{S} = \{(\widetilde{\mathbf{x}}_1, w_{\widetilde{\mathbf{x}}_1}), \ldots, (\widetilde{\mathbf{x}}_m, w_{\widetilde{\mathbf{x}}_{\widetilde{m}}})\}$, where generally $m \neq \widetilde{m}$.

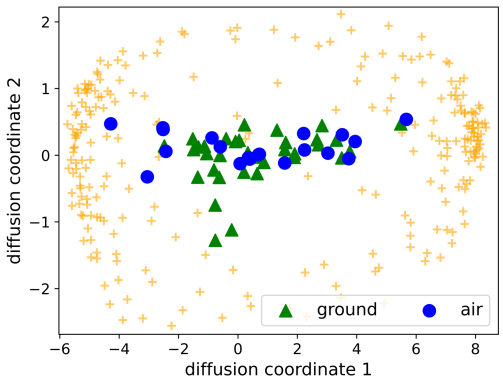

Figure 2: Diffusion map of 54 smoke samples from prescribed burns at BFRS. Each sample (blue circles for air; green triangles for ground) is represented as a convex combination of 250 compound "words" (orange markers); compare with the histogram vs bin arrangement in Figure 1 right. Samples at air versus ground level display differences which are not noticeable when using standard techniques to analyze chemical fingerprint data.

insight as to how differences in smoke type (if observed and statistically significant) relate to groups of compounds. After understanding which fuels, burn conditions, and groups of compounds result in certain negative outcomes like poor health or air quality, we will work with forest managers at BFRS and CAL FIRE to develop a forest management plan that reduces climate-induced wildfire risk while minimizing impacts on the regional environment.

Acknowledgments and Disclosure of Funding

This work is supported in part by the NSF AI Planning Institute for Data-Driven Discovery in Physics, NSF PHY-2020295.

References

[1] S. K. Akagi, R. J. Yokelson, C. Wiedinmyer, M. J. Alvarado, J. S. Reid, T. Karl, J. D. Crounse, and P. O. Wennberg. Emission factors for open and domestic biomass burning for use in atmospheric models. *Atmospheric Chemistry and Physics*, 11(9):4039–4072, 2011.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[3] P. E. Dennison, S. C. Brewer, J. D. Arnold, and M. A. Moritz. Large wildfire trends in the western United States, 1984–2011. *Geophysical Research Letters*, 41(8):2928–2933, 2014.

[4] F. Fouss, A. Pirotte, and M. Saerens. A novel way of computing similarities between nodes of a graph, with application to collaborative recommendation. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 550–556, 2005.

[5] A. H. Goldstein, D. R. Worton, B. J. Williams, S. V. Hering, N. M. Kreisberg, O. Panić, and T. Górecki. Thermal desorption comprehensive two-dimensional gas chromatography for in-situ measurements of organic aerosols. *Journal of Chromatography A*, 1186(1-2):340–347, 2008.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[7] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[8] D. C. Howell. *Statistical methods for psychology*. Cengage Learning, 2009.

[9] C. N. Jen, L. E. Hatch, V. Selimovic, R. J. Yokelson, R. Weber, A. E. Fernandez, N. M. Kreisberg, K. C. Barsanti, and A. H. Goldstein. Speciated and total emission factors of particulate organics from burning western US wildland fuels and their dependence on combustion efficiency. *Atmospheric Chemistry and Physics*, 19(2):1013–1026, 2019.

[10] C. N. Jen, Y. Liang, L. E. Hatch, N. M. Kreisberg, C. Stamatis, K. Kristensen, J. J. Battles, S. L. Stephens, R. A. York, K. C. Barsanti, and A. H. Goldstein. High hydroquinone emissions from burning manzanita. *Environmental Science & Technology Letters*, 5(6):309–314, 2018.

[11] W. M. Jolly, M. A. Cochrane, P. H. Freeborn, Z. A. Holden, T. J. Brown, G. J. Williamson, and D. M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6(1):7537, 2015.

[12] I. Kim, A. B. Lee, and J. Lei. Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305, 2019.

[13] S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.

[14] M. Meila and J. Shi. A random walks view of spectral segmentation. In *Proceedings of AI and Statistics (AISTATS)*, 2001.

[15] M. Roederer and R. R. Hardy. Frequency difference gating: A multivariate method for identifying subsets that differ between samples. *Cytometry*, 45(1):56–64, 2001.

[16] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[17] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.

[18] S. E. Stein and D. R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, 1994.

[19] A. L. Westerling, H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam. Warming and earlier spring increase western U.S. forest wildfire activity. *Science*, 313(5789):940–943, 2006.

[20] C. Wiedinmyer, S. K. Akagi, R. J. Yokelson, L. K. Emmons, J. A. Al–Saadi, J. J. Orlando, and A. J. Soja. The fire inventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning. *Geoscientific Model Development*, 4(3):625–641, 2011.

[21] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205, 2005.