Vehicle Emissions Prediction with Physics-Aware AI Models: Technical Report

Harish Panneer Selvam[1], **Yan Li**[2], **Pengyue Wang,**[1]
William F. Northrop, [1] **Shashi Shekhar**[2]
[1]Dept. of Mechanical Engineering [2]Dept. of Computer Science, University of Minnesota Twin-Cities
Minneapolis, Minnesota 55455
{panne027, lixx4266, wang6609, wnorthro, shekhar}@umn.edu

Abstract

Given an on-board diagnostics (OBD) dataset and a physics-based emissions prediction model, this paper aims to develop an accurate and computational-efficient AI (Artificial Intelligence) method that predicts vehicle emissions values. The problem is of societal importance because vehicular emissions lead to climate change and impact human health. This problem is challenging because the OBD data does not contain enough parameters needed by high-order physics models. Conversely, related work has shown that low-order physics models have poor predictive accuracy when using available OBD data. This paper uses a divergent window co-occurrence pattern detection method to develop a spatiotemporal variability-aware AI model for predicting emission values from the OBD datasets. We conducted a case-study using real-world OBD data from a local public transportation agency. Results show that the proposed AI method has approximately 65% improved predictive accuracy than a non-AI low-order physics model and is approximately 35% more accurate than a baseline model.

1 Introduction

On-board diagnostics (OBD) data is high-resolution multi-attribute trajectory data obtained from various sensors in vehicles. It contains a time series of various engine and vehicle parameters that can be used to model the performance of engine-vehicle systems. Guided by a low-order combustion-physics-based model, this paper aims to develop an OBD-data-driven AI model to predict vehicle emissions values.

The problem of emissions prediction from vehicles is of significant societal importance because transportation is the biggest contributor in the world to greenhouse gases such as CO_2(Carbon dioxide) and toxic gases like NO_x (Oxides of Nitrogen like NO, NO_2, etc.). These emissions lead to over a hundred thousand deaths in the US annually (Thakrar et al. 2020) and also contribute to climate change phenomena such as global warming and acid rain. An understanding of vehicle and powertrain behavior in real-world conditions is essential for tracking and eventually mitigating these emissions by aiding the design of cleaner and more efficient engines and vehicles.

Predicting vehicle emissions values is challenging since the processes by which they are produced are complex and dependent on many parameters. Traditional laboratory experiments conducted to measure emissions values are usually based on engine-specific steady-state measurements. However, data collection inside a laboratory is expensive compared to low-cost sensor data on vehicles already on the road.

Related work in predicting vehicle emissions is of two kinds: purely phenomenological methods or purely AI approaches. (He, Durrett, and Sun 2008) introduced a low-order physics (LOP) model that uses a purely phenomenological method for predicting the emission index of NO_x ($EI{-}NO_x$, grams of NO_x per kilogram fuel consumed) of a diesel engine using vehicle-measured data as input. This method considers the instantaneous amount of fuel burnt during an engine cycle (intake, compression, power, and exhaust strokes) and the instantaneous heat released. The NO_x emissions prediction portion of the model is based on an extension of the Zeldovich mechanism (Mellor et al. 1998) for describing the rate of formation of NO_x. It assumes that $EI - NO_x$ depends on the intake oxygen concentration, residence time of combustion, and the peak adiabatic (i.e. without heat loss) flame temperature. The model is validated using observations from engine testing in laboratory conditions on an apparatus called a dynamometer. We evaluated the LOP method using an OBD dataset (Appendix A) from Metro Transit (a local public transportation agency) buses running three routes in the Minneapolis-St. Paul Region over 5 days (≈16 runs) logged at 1Hz. The dataset contains the measurements of 90 engine and vehicle parameters that can be used to model the working of the *Cummins ISB6.7* Diesel Hybrid engine system of the transit buses in the study. Since the OBD data did not contain certain parameters needed by the LOP method, we estimated the missing parameters from available data as detailed in Appendix B. Figure 1 shows the predicted NO_x emissions compared with the actual values in the dataset. The closer all the dots are to the line of $y = x$, the more accurate the prediction. The absolute values of the prediction errors are color-coded from blue to yellow, where yellow means large prediction errors. Figure 1 shows that the LOP method has poor accuracy when it is used to

predict real-world vehicular NO_x emission values.

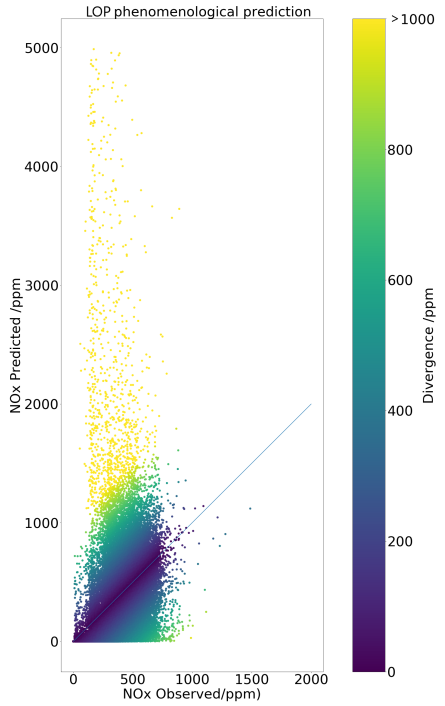


Figure 1: Comparing Observed and Predicted NO_x values using LOP phenomenological model

An example of a purely AI approach is (Obodeh 2009), which evaluates the performance of an artificial neural network (ANN) on data from a laboratory test rig for an engine. However, it provides no understanding of the formation of NO_x and had spurious non-physical results. Instead, engine scientists preferred an approach to predicting emissions that is interpretable using domain knowledge (Karpatne et al. 2018; Karpatne et al. 2017; Li et al. 2020; 2018).

Contributions: In this paper, we propose a novel physics-aware AI model that leverages the concepts of variability across driving scenarios, co-occurrence patterns, and a low-order combustion-physics-based model. We evaluate the proposed model using on-board diagnostics data from Metro Transit (a local public transportation agency) buses. The evaluation results show that the proposed physics-aware AI model predictions are more accurate (65% lower RMSE for training data) than those of the low-order physics model for our OBD dataset.

Scope: The scope of this paper is limited to physics-aware, transparent, and interpretable AI models such as co-occurrence rules guided by a low-order combustion-physics-based model. Other AI models such as neural networks fall outside the scope of this work. Proprietary manufacturers' combustion-physics-based AI models are also outside the scope, due to lack of public availability of proprietary engine calibration data. This paper focuses on the prediction of NO_x emissions from vehicles, thus the prediction of other vehicular emissions is not considered.

Relation to Artificial Intelligence: The 2019 update of the National Artificial Intelligence Research and Development Strategic Plan (Kratsios, Córdova, and Walker 2019) describes *Data Analytics* as one of the main long-term investments that are needed to advance AI. "*Further investigation of multimodality machine learning is needed to enable knowledge discovery from a wide variety of different types of data (e.g., discrete, continuous, text, spatial, temporal, spatio-temporal, graphs).*" Hence, the work presented in this paper is of direct relevance to the AI community.

Outline: The rest of the paper is organized as follows: Section 2 summarizes our baseline method for predicting emissions and Section 3 describes the proposed spatiotemporal variability-aware AI approach to improve the performance. The experimental evaluation is given in Section 4 and Section 5 concludes the paper.

2 Proposed Baseline Approach

We first introduce a baseline physics-aware AI model to predict NO_x emission values. The emission index for NO_x emissions $(EI - NO_x)$ is given by a form of the chemical kinetic equation for the extended Zeldovich Mechanism (Mellor et al. 1998) of formation of NO_x :

$$EI - NO_x(k+\delta) = a * T_{adiab}(k)^b * t_{comb}(k)^c \quad (1)$$

where, $EI - NO_x(k+\delta)$ is the $EI - NO_x$ in grams NO_x per kilogram fuel at time 'k+δ'; $T_{adiab}(k)$ is the adiabatic flame temperature in kelvin at time 'k'; t_{comb} is the duration of combustion in seconds at time stamp 'k', which is approximately equal to the fuel injection duration; a, b, c are constants; and δ is the time lag between the adiabatic flame temperature T_{adiab} and duration of combustion t_{comb} with the corresponding NO_x emission index $EI - NO_x$.

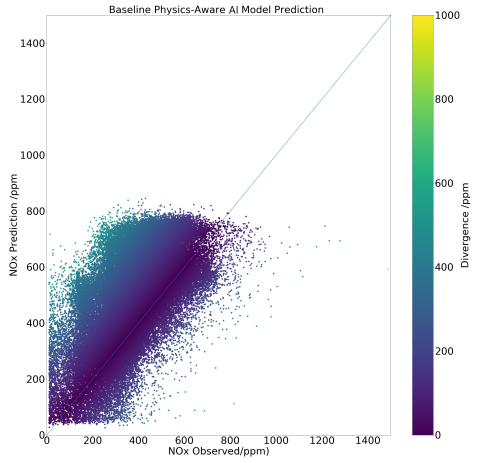


Figure 2: Comparing Observed and Predicted NO_x values using baseline physics-aware AI model

We evaluated the baseline physics-aware AI model using the same OBD dataset as the one used in our initial work for evaluating the LOP method. First, we used six of the observed engine attributes (i.e. intake air flowrate(kilograms

per hour), fuel consumed (kilograms per hour), rail pressure (pascal), intake pressure (pascal), intake temperature (kelvin), engine speed (revolutions per minute)) to calculate T_{adiab} and t_{comb} (Appendix B). Then, we applied a nonlinear regression method provided in the Python Scikit-Learn package (Pedregosa et al. 2011) to estimate the values of a, b and c in Equation 1. The value of $\delta = 1$ was derived via hand computation and data visualization as illustrated in Appendix C.

Figure 2 shows the predicted NO_x values using the baseline model compared with the actual values. The baseline model is an improvement over the low-order physics (LOP) model (Figure 1). However, there is plenty of room for further improvement.

3 Proposed Variability-Aware Approach

To overcome the limitations of the baseline method, we propose a spatiotemporal (ST) variability-aware AI approach. Since one group of estimated parameters (e.g., a, b, c) values in Equation 1 does not fit all scenarios well, it may be beneficial to initially partition the data into multiple homogeneous groups, and estimate parameter values independently for each group.

The top half of Figure 3 shows our proposed ST variability-aware AI framework. First, we test in-coming OBD data to identify NO_x emissions that diverge from the predictions made by the baseline model. We define divergence as the large (i.e. above a given threshold) absolute error between the observed and predicted NO_x values. In general, when a vehicle exhibits divergence, there are two potential pathways for understanding the issues and improving the model: (1) using AI to improve the prediction results, or (2) using physics-based methods to develop new and refined process-based mechanistic models.

This paper focuses on AI model refinement based on data partitioning and fitting separate models to each partition. The partitioning is based on ST correlates of divergent observations (detailed in the next paragraph), thus we name it as an ST variability-aware AI approach. This approach has the potential to reduce prediction errors as illustrated in Figure 3 (lower half).

A divergent window of NO_x emissions refers to a period of a certain length in a time series of OBD data records within which the prediction errors of the baseline approach exceed an input threshold (*summationThreshold*). A co-occurrence pattern in a time series of OBD data records is similar to a sequential association pattern (Srikant and Agrawal 1996) except for the use of spatial statistical interest measure, i.e., Ripley's cross-k function (ϵ) specialized for time series (Ali et al. 2017; Ripley 1976). In other words, the pattern represents those subsets of engine attributes and their specific value ranges, which are present together in many divergent (time-) windows and have cross-k function values above a given threshold (ϵ). More details about the pattern detection algorithm are given in Appendix E. Engine scientists review and group co-occurrence patterns into scenarios (for example, the cold start of an engine, sudden acceleration from a stop, etc.) for ease of interpretation of situations where the baseline model performs poorly. Table

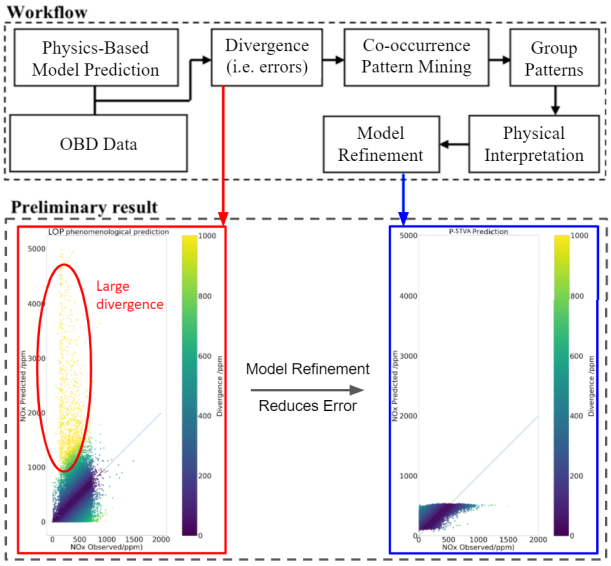


Figure 3: Proposed ST Variability-aware AI framework. The figure in the bottom half is for illustration purpose only.

1 shows examples of co-occurrence patterns from the Metro Transit OBD dataset along with their grouping and interpretation. For example, the group "Low Vehicle Speed Condition" has two examples of co-occurrence patterns based on low wheel speed, represented as w_0 in a time period of three time points where w_0 represents the lowest value range for the wheel speed parameter. The magnitude of each attribute is linearly discretized into 10 equal intervals represented by the subscripts. Similarly, the mined patterns with very low exhaust gas recirculation (EGR) rate and those during transient events are grouped, since the EGR rate is an important factor influencing NO_x formation and transient events cannot be derived from stationary laboratory conditions.

Given the co-occurrence pattern groups formed in the first step, the original OBD dataset is split into multiple subsets corresponding to different pattern groups. Within each subset, we use the baseline approach to calculate the values of T_{adiab} and t_{comb}, and then estimate the parameter (a, b, c) values in Equation 1 by fitting nonlinear regression models independently. Since the many scenarios when the baseline approach does not perform well are handled separately, the ST variability-aware AI model is expected to yield better predictive accuracy by lowering errors.

4 Experimental Evaluation

We conducted experiments to compare the predictive accuracy of the proposed approaches with the low-order physics (LOP) approach detailed in Section 1 (He, Durrett, and Sun 2008) to address the following questions: (1) How do the predictions of the proposed approaches compare with those from the low-order physics approach? (2) How sensitive is the proposed spatiotemporal variability-aware AI approach to the number of partitions, input divergence threshold, and window length?

Table 1: Examples of divergent co-occurrence patterns

Scenario	Example Patterns	
Low Vehicle Speed Condition	1. Wheelspeed w_0 w_0 w_0 RailMPa r_2 r_2 r_2 IntakeT I_9 I_9 I_9	2. Wheelspeed w_0 w_0 w_0 IntakeT I_9 I_9 I_9 Fuelconskgph f_2 f_2 f_2
Low EGR Condition	3. Acceleration a_9 a_9 a_9 EGRkgph g_0 g_0 g_0	4. Bkpwr B_6 B_6 B_6 EGRkgph g_0 g_0 g_0
Transient Condition	5. Wheelspeed w_9 w_{10} w_{10} Bkpwr g_0 g_0 g_0	6. Acceleration a_8 a_8 a_7 RailMPa r_6 r_6 r_6

Subscript	Scale of values
0,1	Very low value
2,3,4	Low value
5,6,7	Medium Value
8,9,10	High value

The experiment design is summarized in Figure 4.

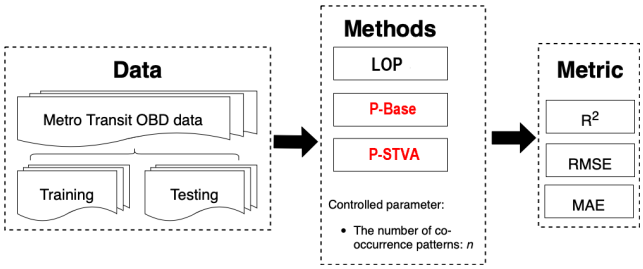


Figure 4: Experimental design

Data: The dataset used in the experiments is the Metro Transit OBD dataset that was used to evaluate the LOP approach and the proposed baseline approach in the earlier sections. It contains 99,895 data entries containing measurements of 90 engine and vehicle attributes (detailed in Appendix A). The OBD data was obtained from transit buses traversing 3 different routes for 16 different runs in the Minneapolis-St.Paul region. We used 8 runs for training and the remaining 8 runs for testing, ensuring that each route was represented in both training and testing samples.

Candidate methods: The methods evaluated in the experiments include low-order physics model (LOP), the proposed baseline (P-Base), and the proposed spatiotemporal variability-aware AI approach (P-STVA).

Metrics: The predictive accuracy was measured using R^2 values, root mean square error (RMSE) and mean absolute error (MAE).

Experimental Results: Figure 5 shows a comparison of the refined NO_x prediction using the P-STVA method with the observed NO_x values in the training data. Compared with Figure 1 and 2, the dots in Figure 5 are closer to the $y = x$ line, and the number of green and yellow dots in the upper-left part of Figure 2 reduces dramatically, which indicates improved predictive accuracy (detailed in Appendix D).

How do the predictions of the proposed approaches compare with those from the low-order physics approach? Table 2 and 3 summarize predictive accuracy metrics for the candidate methods on training and testing data respectively with $n = 4$ and *summationThreshold*=30 ppm (parts per million). The proposed physics-aware AI methods outperformed the low-order physics model. For the training data, the P-base method provides about 50% improvement in RMSE and 35% improvement in MAE when compared to the LOP method, while RMSE and MAE of the P-STVA method with $n = 4$ are both around 35% smaller than the P-base method. For the testing data, the P-base method provides about 50% improvement in RMSE and 40% improvement in MAE when compared to the LOP method, while RMSE and MAE of the P-STVA method with $n = 4$ are both around 35% smaller than the P-base method.

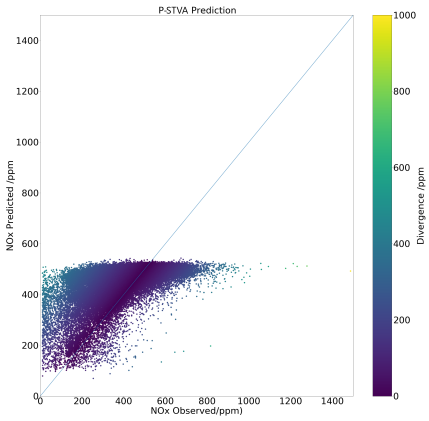


Figure 5: Refined NO_x prediction using P-STVA method, number of partitions = 4, L = 3 seconds, *summationThreshold* = 30 ppm, *minSupp* = 0.003, $\epsilon = 2$

Table 2: NO_x predictive accuracy for training data

Prediction method	R^2	RMSE	MAE
LOP	0.1264	371.16	238.87
P-Base	0.4464	196.39	155.17
P-STVA n=4	0.3900	132.60	102.13

Table 3: NO_x predictive accuracy for testing data

Prediction method	R^2	RMSE	MAE
LOP	0.1260	368.67	238.23
P-Base	0.4607	183.52	144.69
P-STVA n=4	0.4769	117.39	92.99

How sensitive is the proposed spatiotemporal variability-aware AI approach to the number of partitions, input divergence threshold, and window length?

Figure 6 shows the sensitivity of predictive accuracy of the P-STVA method to number of patterns n (number of partitions is $n+1$) on training and testing data with *summationThreshold* = 30 ppm and window length L = 3 seconds. The predictive accuracy improves with increasing number of patterns.

Figure 7 shows the sensitivity of the predictive accuracy to *summationThreshold* for the P-STVA method for training and testing data with $n = 4$ and L = 3 seconds. The RMSE and MAE improve till *summationThreshold* = 30 ppm, and

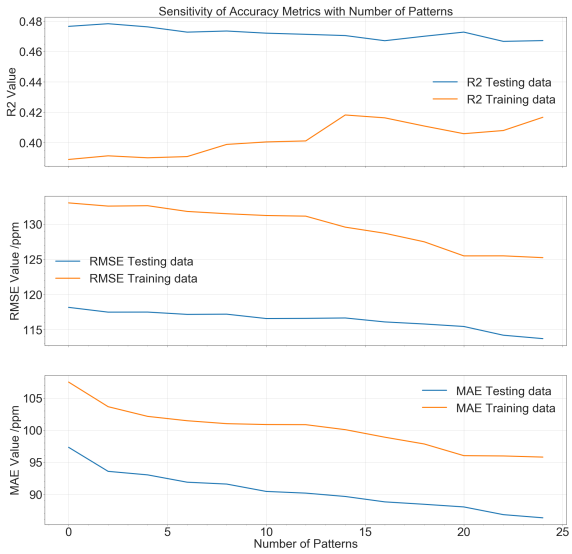


Figure 6: Sensitivity of the predictive accuracy metrics of P-STVA method to the number of co-occurrence patterns used to split the dataset

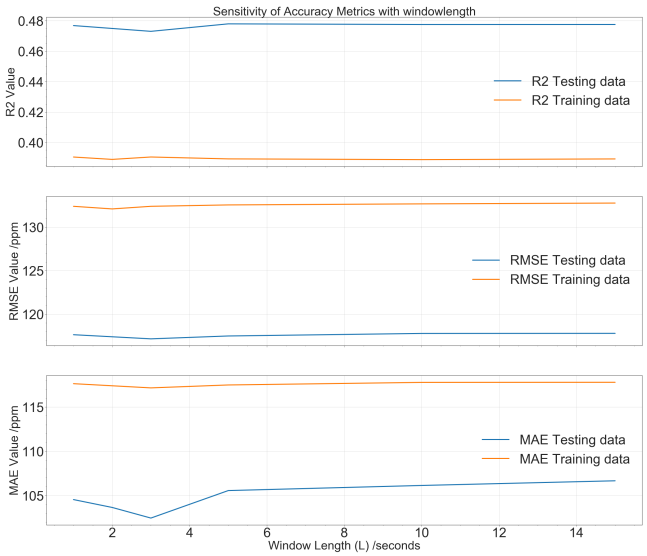


Figure 8: Sensitivity of predictive accuracy metrics of P-STVA method to Window Length L

then diminish.

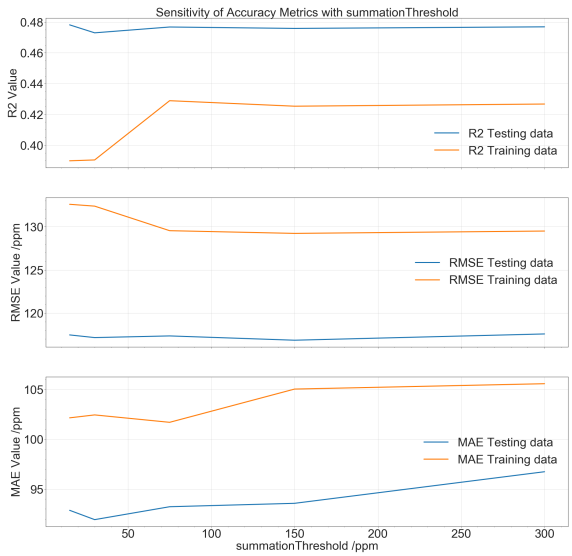


Figure 7: Sensitivity of predictive accuracy metrics of P-STVA method to *summationThreshold*

Figure 8 shows the sensitivity of the predictive accuracy to window length L for P-STVA method for training and testing data with $n = 4$ and *summationThreshold* = 30 ppm. The RMSE and MAE improve till L = 3 seconds, and then diminish.

5 Discussion

Domain interpretation of partitions: The P-STVA method uses $n + 1$ partitions including the non-divergent case (and divergent cases not covered by selected patterns) and additional n partitions of divergent cases, one for each co-occurrence pattern. The top half of Figure 9 shows 4 co-occurrence patterns for the case $n = 4$, *summationThreshold* = 30 ppm and L = 3 seconds along with their domain interpretation in terms of different scenarios. The bottom half of Figure 9 shows the performance of the P-STVA method within each pattern of divergent cases. The short time windows make it difficult to classify the scenarios and patterns into either transient or steady-state. Table 4 in Appendix D shows the values of parameters a, b, and c for four partitions defined by divergence patterns.

Computation time comparison: The proposed P-STVA method took a few minutes of computation time on a Microsoft Surface Pro 7 tablet with Intel i5 CPU and 8GB memory with 99,895 points (80MB) on-board diagnostics data, including the time for mining co-occurrence patterns and estimating parameters (a, b, c) for each partition for $n = 4$, *summationThreshold* = 30 ppm, and L = 3 seconds. Figure 10 shows that computation time decreases with increase in *summationThreshold*, since the number of divergent windows to analyze also decreases. The low-order physics model took substantially more computation time as it had to solve discretized partial differential equations over an engine cycle for each data entry.

Broader literature review: Since proprietary engine calibration data is not publicly available, it is difficult to make a direct comparison between those and our work. Nonetheless, we provide a summary of broader literature for interested users. If engine calibration data is available, it can be used to obtain a brake-specific NO_x value for engine torque and engine speed from OBD data (Rosero et al. 2020). (Kotz, Kittelson, and Northrop 2016) introduced a new method for identifying NO_x emissions hotspots along a bus route using high fidelity Lagrangian vehicle data to explore spatial inter-

Pattern	Co-occurrence	Scenario
Pattern1	EngTq: T_{10} T_{10} T_{10}	High Engine Load
Pattern2	EngTq: T_9 T_{10} T_{10}	High Load Transient
Pattern3	EngRPM: R_1 R_2 R_2 EGRkgph: g_4 g_4 g_4 EngTq: T_1 T_1 T_1	High Engine Idling
Pattern4	EngRPM: R_1 R_2 R_2 EGRkgph: g_4 g_4 g_4	Low Engine Speed

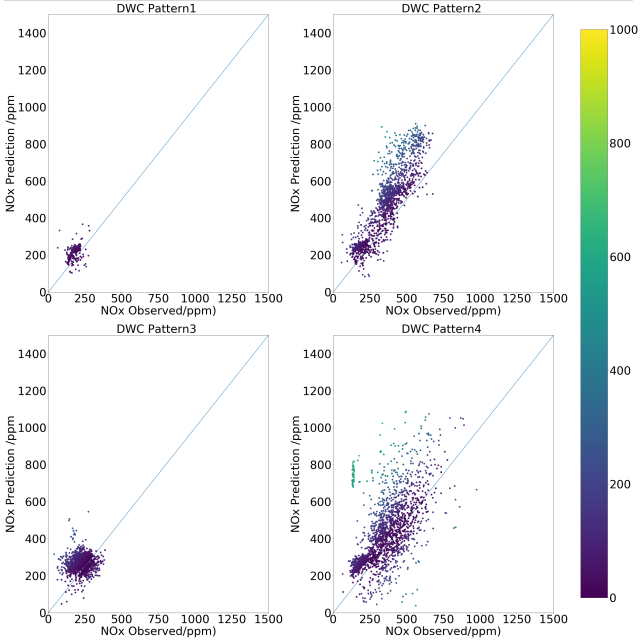


Figure 9: NO_xpredictions of four partitions corresponding to four most significant divergent co-occurrence patterns in the training data

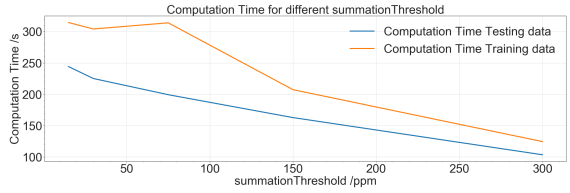


Figure 10: Change in Computation time for change in *summationThreshold* ($n = 4$, *L*=3s)

actions that may influence emissions production. The paper noted that NO_x hotspots occurred at bus stops, during cold starts, on inclines, and for accelerations.

(Jahangiri and Rakha 2015) adopt different supervised learning methods from the field of machine learning to develop multi-class classifiers that identify the transportation mode. The methods which are evaluated include K-nearest neighbor, support vector machines (SVMs), and tree-based models that comprise a single decision tree, bagging, and random forest (RF) methods. Among them, SVM and RF produce comparatively better performances. (Hagenauer and Helbich 2017) did similar work and the authors made a comparison of 7 classifiers for travel mode prediction. (Omrani 2015) presents four machine learning methods for predicting the travel mode of individuals. All these papers are focused on clustering transportation data into different travel modes.

This paper is based on preliminary work by (Ali et al. 2017) on mining co-occurrence patterns from OBD data where the emissions values in a route were not compliant with EPA regulations. The detected patterns were interpreted by domain scientists as different vehicle scenarios.

6 Conclusions and Future Work

We proposed a novel physics-aware AI emission prediction model and evaluated it with an on-board diagnostics dataset. The experimental evaluation shows the proposed models outperform the non-AI low-order physics model. Furthermore, the resultant models were interpreted using domain concepts as different vehicle scenarios.

In the future, we will explore other AI models such as neural networks guided by combustion physics. We will characterize the sensitivity of the computation time of the proposed P-STVA method to parameters such as the number of partitions. We will also investigate physics-aware AI models to predict vehicle emissions other than NO_x.

7 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants No. 1029711, 1737633, IIS-1320580, IIS-0940818, IIS-1218168, 1916518, and 1901099, the USDOD under Grants No. HM1582-08-1-0017 and HM0210-13-1-0005, the Advanced Research ProjectsAgency-Energy (ARPA-E), USDOE under Award No. DE-AR0000795, the NIH under Grant No. UL1 TR002494, KL2TR002492, and TL1 TR002493, the USDA under Grant No. 2017-51181-27222. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

Ali, R.; Gunturi, V.; Kotz, A.; Eftelioglu, E.; Shekhar, S.; and Northrop, W. 2017. Discovering non-compliant window co-occurrence patterns. *GeoInformatica* 21(4):829–866. Springer,US.

Gabriel, E., and Diggle, P. J. 2009. Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica* 63(1):43–51. Wiley Online Library.

Hagenauer, J., and Helbich, M. 2017. A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Syst. Appl.* 78(C):273–282. Pergamon Press, Inc.

He, X.; Durrett, R. P.; and Sun, Z. 2008. Late intake valve closing as an emissions control strategy at tier 2 bin 5 engine-out nox level. *SAE Int. J. Engines* 1:427–443. SAE International.

Heywood, J. B. 2018. *Internal combustion engine fundamentals*. New York ; London: McGraw-Hill Education, second edition.

Jahangiri, A., and Rakha, H. A. 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE Transactions on Intelligent Transportation Systems* 16(5):2406–2417. IEEE.

Karpatne, A.; Atluri, G.; Faghmous, J. H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; and Kumar, V. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering* 29(10):2318–2331. IEEE.

Karpatne, A.; Kumar, V.; Jia, X.; Read, J.; and Hanson, P. 2018. Physics guided machine learning: a new paradigm for modeling dynamical systems. *AGUFM* 2018:IN12A–03. AGU.org.

Kotz, A. J.; Kittelson, D. B.; and Northrop, W. F. 2016. Lagrangian hotspots of in-use nox emissions from transit buses. *Environmental Sc. & Tech.* 50(11):5750–5756. American Chemical Society.

Kratsios, M.; Córdova, F. A.; and Walker, S. 2019. National artificial intelligence research and development strategic plan: 2019 update. *National Science & Technology Council, USA*.

Li, Y.; Shekhar, S.; Wang, P.; and Northrop, W. 2018. Physics-guided energy-efficient path selection: a summary of results. In *SIGSPATIAL'18*, 99–108. ACM.

Li, Y.; Kotwal, P.; Wang, P.; Shekhar, S.; and Northrop, W. 2020. Physics-guided energy-efficient path selection using on-board diagnostics data. *ACM Transactions on Data Science (TDS)* 1(1):. ACM.

Mellor, A. M.; Mello, J. P.; Duffy, K. P.; Easley, W. L.; and Faulkner, J. C. 1998. Skeletal mechanism for nox chemistry in diesel engines. *SAE Transactions* 107:786–801. SAE International.

Obodeh, O. 2009. Evaluation of artificial neural network performance in predicting diesel engine nox emissions. *Research Journal of Applied Sciences, Engineering and Technology* 33. Maxwell Scientific Publications.

Omrani, H. 2015. Predicting travel mode of individuals by machine learning. *Transportation Research Procedia* 10:840–849. Elsevier B.V.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(85):2825–2830. JMLR.org.

Ripley, B. D. 1976. The second-order analysis of stationary point processes. *Journal of applied probability* 13(2):255–266. Cambridge University Press.

Rosero, F.; Fonseca, N.; López, J.-M.; and Casanova, J. 2020. Real-world fuel efficiency and emissions from an urban diesel bus engine under transient operating conditions. *Applied Energy* 261. Elsevier Ltd.

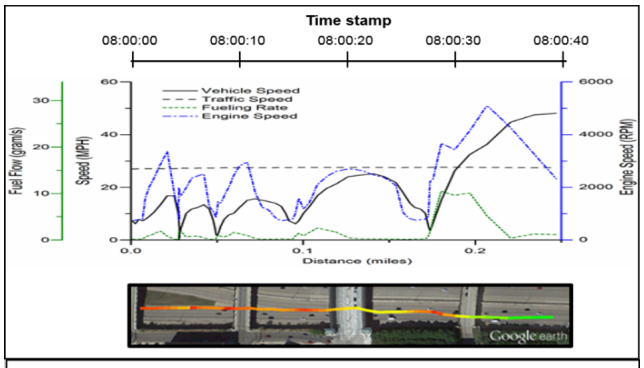
Srikant, R., and Agrawal, R. 1996. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, 3—-17. Springer–Verlag. Berlin, Heidelberg: Springer-Verlag.

Thakrar, S.; Balasubramanian, S.; Adams, P.; Azevedo, I.; Muller, N.; Pandis, S.; Polasky, S.; Pope, C.; Robinson, A.; Apte, J.; Tessum, C.; Marshall, J.; and Hill, J. 2020. Reducing mortality from air pollution in the united states by targeting specific emission sources. *Environmental Science and Technology Letters.* American Chemical Society.

Turns, S. R. 2000. *An Introduction to Combustion : Concepts and Applications*. McGraw-Hill series in mechanical engineering. Boston: WCB/McGraw-Hill, 2nd edition.

Appendix A On-Board Diagnostics Data

On-board diagnostics (OBD) data is a high-resolution multi-attribute trajectory data obtained from sensors in modern vehicles. Recorded parameters in the dataset have 1 Hz resolution consisting of around 90 columns including time, GPS location, ambient conditions, emissions, and engine operating parameters among others. A sample dataset, collected in the Minneapolis - St. Paul area, is shown in Figure 11. The red-colored areas denote route segments with excessive NO_x emissions.



Vehicle id	Time stamp	Location (lat, long)	Map-matched edge-id	Attributes		
				NO_x	Engine RPM	. . .
V1	08:00:00	44.9778° N, 93.2652° W	E1	0.492	860	
V1	08:00:10	44.9778° N, 93.2650° W	E2	0.492	2,980	
V1	08:00:20	44.9777° N, 93.2649° W	E3	0.532	2,730	
V1	08:00:30	44.9775° N, 93.2648° W	E4	0.352	3,870	
V1	08:00:40	44.9774° N, 93.2647° W	E5	0.452	2,321	
. . .						

Figure 11: OBD Data and its tabular representation

Appendix B Physics Calculations

Calculating Adiabatic Flame Temperature:
Data Used: *Intake Air Flowrate (kilograms per hour), Fuel consumed (kilograms per hour)*

The Adiabatic Flame Temperature (constant volume) for the combustion reaction of diesel inside the engine chamber is the temperature of the products of the reaction when there is no shaft work done. An energy balance equating the net calorific value of the fuel consumed with the temperature-dependent specific heats (Turns 2000), and compositions of the different products that depend on the equivalence ratio, was used to compute the Adiabatic Flame Temperature.

Calculating Duration of Combustion:

Data Used: *Rail Pressure (pascal), Intake Pressure (pascal), Intake Temperature (kelvin), Fuel consumed(kilograms per hour), Engine speed (revolutions per minute)*

The duration of combustion or residence time for the diesel fuel inside a combustion chamber is approximately equal to the fuel injection duration. The fuel injection duration was calculated from the fluid flow equation (Heywood 2018) of the *Cummins ISB6.7* injector nozzles assuming a flow coefficient of 0.86.

Calculating EI-NO_x (observed) :

Data Used: *Fuel consumed (kilograms per hour), SCR(Selective Catalytic Reduction) flowrate (grams per second)*

$EI - NO_x$ (observed) is Emission Index of NO_x (in grams of NO_x per kilogram of fuel consumed) that is used as the prediction target in curve fitting the features (T_{Adiab}, t_{comb}) to obtain a prediction of $EI - NO_x$ values. $EI - NO_x$ is converted to ppm (parts per million) by calculating the number of moles of NO_x and dividing by the total number of moles of product for a given timestamp.

Appendix C Data Visualization

Different subsets of the OBD dataset were selected based on wheel speed data. Sixty-seconds long subsets like in Figure 12 corresponding to different vehicle scenarios such as accelerating(increase in wheel speed), coasting (constant wheel speed), braking (decrease in wheel speed), etc., were selected. This helped visualize the correlations between the observed NO_x values, the duration of combustion, and the adiabatic flame temperature. It is also observed that the measured NO_x occurs with a lead of 1 second, which is accounted for through δ in Equation 1

Appendix D P-STVA Model Parameters

Table 4 shows the sets of parameters from Equation 1 corresponding to four divergent window co-occurrence patterns. These patterns were detected from the OBD dataset in Figure 9 using the proposed ST variability-aware AI method. The parameters [a, b, c] obtained from the proposed baseline method for training data are [4.86e+09, -2.15686744e+00, 5.33873453e-01]

Table 4: P-STVA method parameters [a,b,c] in Equation 1

Parameter	Pattern1	Pattern2	Pattern3	Pattern4
a	1.39E+09	5.64E+09	2.46E+10	1.49E+07
b	-1.98802	-2.10515	-2.12836	-1.57837
c	0.505761	0.593693	0.839214	0.256813

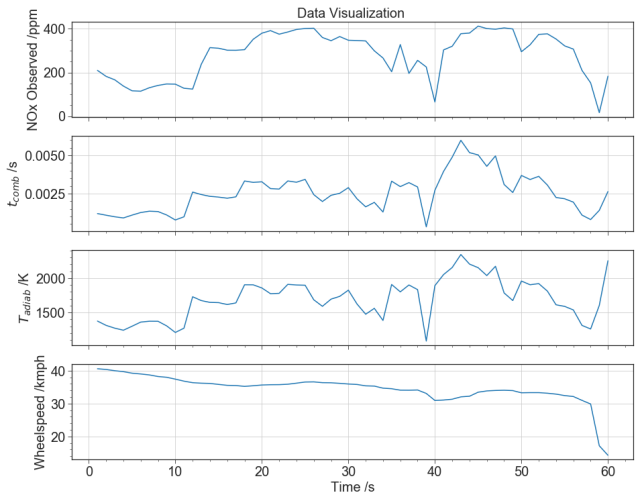


Figure 12: Correlation between NO_x Observed, Adiabatic Flame Temperature, Duration of Combustion for sixty-seconds long subset where the vehicle is coasting (constant speed) and then braking

Appendix E DWC Pattern Detection

Given a collection of multi-attribute trajectories (m) defined over a set of explanatory variables and a target variable, and a time lag, a divergent window co-occurrence (DWC) pattern refers to a collection of equal-length time sequences of a subset of the input explanatory variables that start at the same time point, and within the input time lag preceding the start of a divergent window. A divergent window is a time interval of a given length within which the target variable meets certain criteria. Suppose that time windows wherein predicted NO_x emissions value being divergent from observed value is of interest, divergent window co-occurrence pattern detection can find collections of engine measurement value sequences that tend to occur together within the divergent windows.

The temporal cross-K function is used to express how much the association between a given pattern and divergent windows deviates from independence (Ali et al. 2017) and a minimum support (i.e. frequency) threshold is used to filter out chance patterns that rarely occur. The temporal cross-K function is a purely temporal form of the space-time cross-K function (Gabriel and Diggle 2009).