# Machine Learning for Activity-Based Road Transportation Emissions Estimation

**Derek Rollend** **Kevin Foster** **Tomek Kott** **Rohita Mocharla** **Rai Muñoz**
**Neil Fendley** **Chace Ashcraft** **Frank Willard** **Marisa Hughes**
The Johns Hopkins University Applied Physics Laboratory
{firstname}.{lastname}@jhuapl.edu

## Abstract

Measuring and attributing greenhouse gas (GHG) emissions remains a challenging problem as the world strives towards meeting emissions reductions targets. As a significant portion of total global emissions, the road transportation sector represents an enormous challenge for estimating and tracking emissions at a global scale. To meet this challenge, we have developed a hybrid approach for estimating road transportation emissions that combines the strengths of machine learning and satellite imagery with localized emissions factors data to create an accurate, globally scalable, and easily configurable GHG monitoring framework.

## 1 Introduction

Transportation contributed 27% of anthropogenic greenhouse gas (GHG) emissions in the U.S. for 2020, higher than any other sector, and 12.6% of all global GHG emissions in 2019 [1, 2]. The primary source of transportation sector emissions are on-road vehicles, accounting for approximately 74% of global transportation emissions in 2018 [3]. Quantifying the distribution of on-road transportation emissions and creating timely emissions inventories are vital to identify trends, track mitigation efforts, and inform policy decisions.

Previous efforts have developed detailed bottom-up on-road emission inventories for the U.S. [4, 5], but do not easily extend globally due to the reliance on vehicle traffic and road data that is not always readily available. EDGAR [6] provides a global inventory for transportation that uses road density as a proxy to spatially distribute emissions. However, some emission estimates for urban centers in EDGAR deviated from other bottom-up inventories [4] by 500%, indicating that road density is not a sufficient proxy for global high-resolution inventories. Carbon Monitor [7] is a global emissions inventory that utilizes a variety of activity data to estimate daily GHG emissions, however the reliance on proprietary traffic data in the ground transportation sector limits the ability to extend to locations where this data is not available. Other methods have used machine learning (ML) to directly predict emissions, but their ability to generalize globally is unclear [8, 9].

We propose an emissions estimation method that is globally accurate while using openly available input data. It combines remote sensing, geospatial data, and ML with traditional, “bottom-up" emissions inventories that directly incorporate region-specific vehicle fleet mix, fuel efficiency, and other emissions factors (EF) data. This approach, illustrated in Figure 1, is composed of ML models to predict road transportation activity and an EF pipeline that translates activity to emissions in a localized fashion. These two independent parts afford continuous improvement as newer data become available. Our contribution is a method that uses ML-predicted road activity along with

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

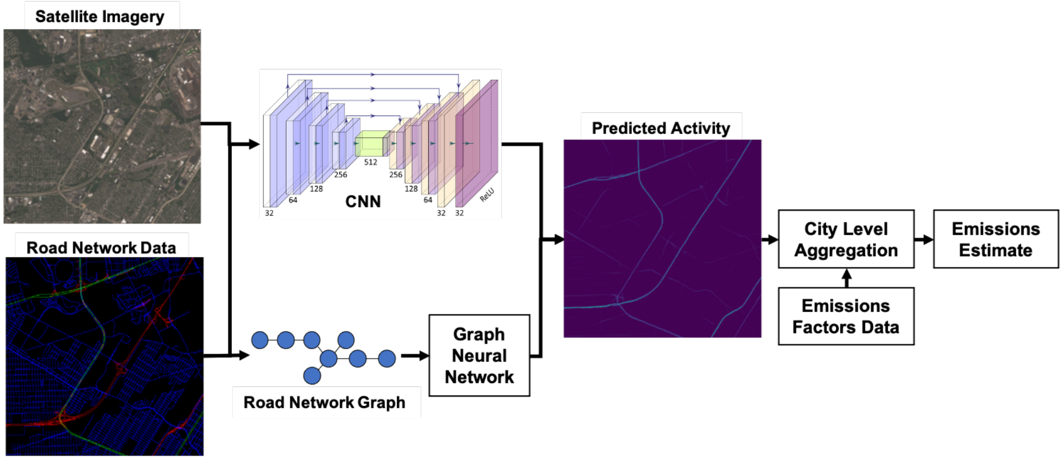

Figure 1: System architecture for our hybrid emissions estimation model.

region-specific emissions factors data to create up-to-date and more accurate global on-road GHG emissions estimates.

## 2 Data & Methods

We formulate the road activity prediction task as a regression problem. We train models to predict average annual daily traffic (AADT), or the number of vehicles traveling on a given road segment per day, on average over an entire year. We use ground truth data from the U.S. Highway Performance Monitoring System Average Annual Daily Traffic (AADT) data set from 2017 [10]. This AADT data is recorded using road-side devices, and is typically only recorded on major highways and arterial (collector) roads. Models trained in the U.S. are then run over both U.S. and global cities for evaluation.

OpenStreetMap (OSM) [11] road network data is used as input to our models and for associating predicted AADT values with their corresponding physical road segment. While OSM can contain inconsistencies and is not complete, its open access and global availability make it suitable for this work.

**Convolutional Neural Networks** Semantic segmentation convolutional neural networks (CNNs) were trained to predict AADT, using visual RGB satellite imagery and road network data. Separate models were trained using two sources of imagery: Sentinel-2 Level-2A visual RGB at 10 m x 10 m resolution [12], and Planet Labs PlanetScope mosaics at ∼3 m resolution [13]. OSM data is rasterized for the corresponding extent of an input visual image, where each road type (highway, secondary, local) is rasterized independently, and the resulting raster channels are concatenated together to form a three channel image (see Appendix A for associated OSM tags for each road type). This image is combined with the visual image to form a six channel image that is fed to the CNN to predict AADT on a per-pixel basis. We primarily use MANet-based architectures [14] for our segmentation models, with EfficientNet [15] backbones (see Appendix B for full model descriptions). All AADT predictions for a given road segment are averaged to produce a single AADT value for every road within the current geographic extent of the input data.

**Graph Neural Networks** We have also trained graph neural networks (GNNs) [16] to predict AADT. Road networks inherently take the form of a graph structure, and a graph neural network (GNN) can capture road activity and feature dependencies across a range of scales more easily than the image-based CNN segmentation models. GNNs can easily leverage various features assigned to nodes and efficiently reason over the full road network graph to provide more robust estimates of on-road activity. A number of road features are derived from OSM for model input, including: road length, road type, number of lanes, and the directional angle between roads. The graph attention network v2 (GATv2) [17] architecture is used as it allows for both edge and node input features, and is set up to predict log-AADT values. Further model details can be found in Appendix B.

**Model Ensembling** To create a more robust and predictive AADT estimation model, ensembling is performed using the CNN and GNN models. Model AADT predictions per road segment are averaged before being input to the emissions factors pipeline. This capability can be easily extended in the future to experiment with different model architectures and perform further analysis of inter-model variance. An example ensemble AADT output can be seen in Figure 2.

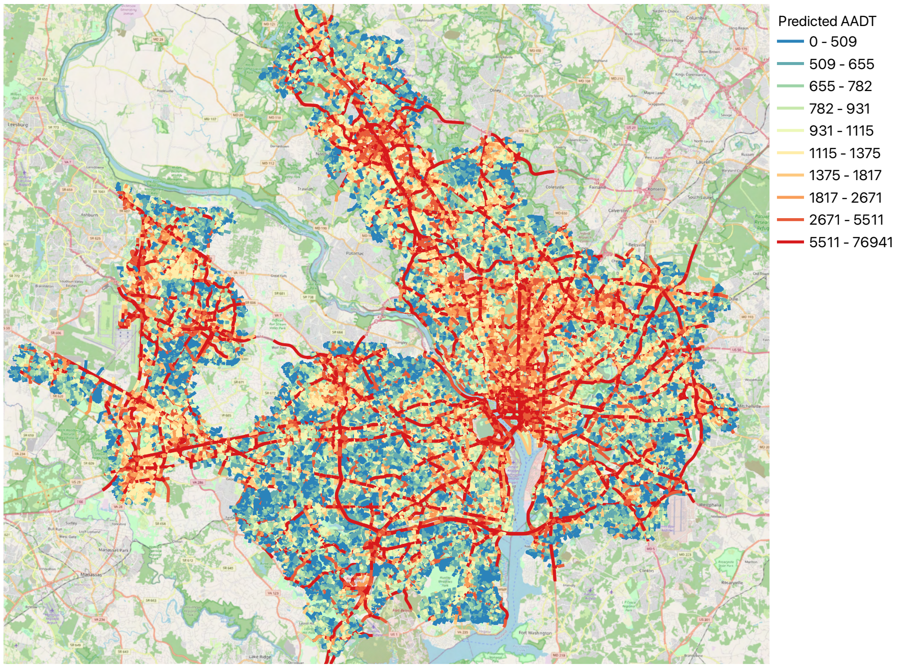

Figure 2: Example ensembled AADT predictions for the greater Washington, D.C. area. AADT units are vehicles per day. Map data from OpenStreetMap [11].

**Activity to Emissions** AADT predictions are assigned to their corresponding road segment based on the known geographic location of the underlying road network. Emissions factors are computed *a priori* from a database of road and vehicle-related data for a specific region, assigning EF values to each type of road in a city. Estimated AADT is multiplied by 365 and converted to total annual vehicle kilometers traveled (VKT) using the known length of each road segment, and then multiplied by the corresponding EF for each road type. This process is repeated and summed over all road segments in a city to calculate the final, total annual GHG emissions estimates for that city (see Appendix C for more information). To date, our hybrid emissions estimation pipeline has been run on a prioritized set of 500 global cities (see Appendix D).

## 3 Results & Discussion

**U.S. Activity Prediction** We evaluate each ML model on a hold out test set of U.S. cities, using input data from 2017 to align with the timespan of our ground truth AADT data. We compute the following metrics on a per road basis: root mean squared error (RMSE), mean absolute percentage error (MAPE), mean percentage error (MPE), and Pearson's $\rho$ (see Appendix E for definitions). Comparing metrics on a per-road basis enables a fair comparison between the image-based convolutional neural network (CNN) models and the graph-based GNN models, as can be seen in Table 1. Of note is the lower error metrics for the CNN models, but the stronger correlation of the GNN models. This points to the importance of model ensembling to create a more robust activity prediction.

**International Activity Prediction** To estimate the ability of our models to generalize outside of the U.S.-based training set, we have run an ensemble of the S2+OSM CNN and GNN OSM models on 500 global cities, using input data from 2021. Several international AADT datasets are used for evaluation: 26 cities in the United Kingdom (U.K.) for 2018-2020 (U.K. 26, see Appendix F) [18], Buenos Aires, Argentina [19], and Paris, France [20]. Per-road AADT error metrics between our

Table 1: Comparison of activity prediction models trained with varying inputs and architectures (see Appendix B for a full description of each model). RMSE is in units of vehicles per day.

| Method | RMSE | MAPE | MPE | Pearson's $\rho$ |
|---|---|---|---|---|
| S2+OSM | 4823.6 | 116.3% | **-39.3%** | 0.58 |
| S2+OSM Ensemble | 5249.5 | **102.3%** | -71.74% | 0.58 |
| Planet+OSM | **3329.9** | 159.9% | 41.01% | 0.60 |
| GNN OSM | 4470.0 | 137.6% | 103.3% | 0.87 |
| GNN OSM+GHSL | 4384.9 | 143.3% | 110.6% | 0.87 |
| GNN OSM+CNN | 4415.3 | 135.0% | 99.27% | **0.88** |
| GNN OSM Ensemble | 4307.3 | 142.3% | 113.0% | **0.88** |

ML estimates and these datasets are shown in Table 2. Error percentages are generally on par with performance in the U.S., showing the ability of our models to generalize globally.

Table 2: Evaluation of ensembled model output with international AADT.

| Region | RMSE | MAPE | MPE | Pearson's $\rho$ |
|---|---|---|---|---|
| U.K. 26 (2018) | 3804.6 | 119.5% | 49.2% | 0.69 |
| U.K. 26 (2019) | 3177.2 | 130.9% | 63.0% | 0.73 |
| U.K. 26 (2020) | 3447.0 | 84.7% | 20.6% | 0.69 |
| Buenos Aires (2017) | 8750.2 | 74.3% | 71.8% | 0.66 |
| Paris (2021) | 9467.9 | 96.3% | 20.4% | 0.79 |

**Emissions Validation** We have performed several comparisons of our road transportation emissions estimates against other emissions inventories for initial validation, both within the US and globally. A set of 14 hold out cities in the US were selected for validation with three other emissions inventories: Google Environmental Insights Explorer (EIE) [21], Database of Road Transportation Emissions (DARTE) [22], and Vulcan v3.0 [5]. Both our CNN and GNN-derived emissions estimates are strongly correlated with other inventory values for every city, with mean Pearson $\rho$ values of 0.97 (CNN) and 0.98 (GNN). We also compare our emissions estimates for 500 of the largest global cities to EDGAR [6] and Carbon Monitor [7]. We found strong correlation with both inventories, with Pearson $\rho$ values of 0.74 and 0.87, respectively, showing the high global accuracy of our method. Further emissions validation details can be found in Appendix G and H.

## 4 Conclusion

We have presented a hybrid road transportation emissions estimation method that is accurate, scalable, and easy to update. The ability to calculate emissions per road segment can be further refined to reach an unprecedented level of detail and global coverage. Where available, the integration of real-time traffic data would increase the temporal resolution and accuracy of our models. We also plan to carry out further analysis of our emissions estimates with other inventories to identify the main causes of discrepancies. As well, we aim to explore open-sourcing our emissions factors schema such that governments and other entities can contribute more up-to-date and accurate EF data to further improve our estimates. This type of actionable emissions monitoring data will be critical to ensuring we meet global emissions reduction targets and may inspire new ways of mitigating the effects of climate change.

## Acknowledgments and Disclosure of Funding

We would like to thank the Climate TRACE coalition, WattTime, Google.org, Generation Investment Management, and Al Gore for their organizational and financial support. Additional thanks to Dr. Kevin Gurney, Northern Arizona University, for his continuous feedback and review of our methodology.

## References

[1] *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2020*. Tech. rep. U.S. Environmental Protection Agency, 2022.

[2] World Resource Institute. *Climate Watch Historical GHG Emissions*. https://www.climatewatchdata.org/ghg-emissions. 2022.

[3] International Energy Agency (IEA). *Transport sector CO2 emissions by mode in the Sustainable Development Scenario, 2000-2030*. URL: https://www.iea.org/data-and-statistics/charts/transport-sector-co2-emissions-by-mode-in-the-sustainable-development-scenario-2000-2030 (visited on 09/16/2022).

[4] Conor K Gately, Lucy R Hutyra, and Ian Sue Wing. "Cities, traffic, and $CO_2$: A multidecadal assessment of trends, drivers, and scaling relationships". In: *Proceedings of the National Academy of Sciences* (2015).

[5] Kevin R Gurney et al. "The Vulcan version 3.0 high-resolution fossil fuel $CO_2$ emissions for the United States". In: *Journal of Geophysical Research: Atmospheres* (2020).

[6] Greet Janssens-Maenhout et al. "EDGAR v4.3.2 Global Atlas of the three major Greenhouse Gas Emissions for the period 1970–2012". In: *Earth System Science Data Discussions* (2017).

[7] Zhu Liu et al. "Carbon Monitor, a near-real-time daily dataset of global CO2 emission from fossil fuel and cement production". In: *Scientific Data* 7 (Nov. 2020). DOI: 10.1038/s41597-020-00708-7.

[8] Linus M. Scheibenreif, Michael Mommert, and Damian Borth. "Estimation of Air Pollution with Remote Sensing Data: Revealing Greenhouse Gas Emissions from Space". In: *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021. URL: https://www.climatechange.ai/papers/icml2021/23.

[9] Ryan Mukherjee et al. "Towards Indirect Top-Down Road Transport Emissions Estimation". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2021, pp. 1092–1101. DOI: 10.1109/CVPRW53098.2021.00120.

[10] US Federal Highway Administration. *Highway Performance Monitoring System (HPMS) Data*. 2017. URL: https://www.fhwa.dot.gov/policyinformation/hpms.

[11] Mordechai Haklay and Patrick Weber. "OpenStreetMap: User-Generated Street Maps". In: *IEEE Pervasive Computing* (2008).

[12] Matthias Drusch et al. "Sentinel-2: ESA's optical high-resolution mission for GMES operational services". In: *Remote sensing of Environment* (2012).

[13] Planet Labs Inc. *Planet imagery product specifications*. 2022. URL: https://assets.planet.com/docs/Planet_Combined_Imagery_Product_Specs_letter_screen.pdf (visited on 09/13/2022).

[14] Tongle Fan et al. "MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation". In: *IEEE Access* (2020).

[15] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *ICML*. 2019.

[16] Michael M. Bronstein et al. "Geometric Deep Learning: Going beyond Euclidean data". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42. DOI: 10.1109/MSP.2017.2693418.

[17] Shaked Brody, Uri Alon, and Eran Yahav. *How Attentive are Graph Attention Networks?* 2021. DOI: 10.48550/ARXIV.2105.14491. URL: https://arxiv.org/abs/2105.14491.

[18] Department for Transport. *Road Traffic Statistics*. 2020. URL: https://roadtraffic.dft.gov.uk.

[19] Ministry of Transport - National Directorate of Roads. *2017 TMDA data*. 2017. URL: https://datos.transporte.gob.ar/dataset/tmda.

[20] Department of Roads and Travel - Service des Déplacements - Poste Central d'Exploitation Lutèce. *Road counting - Traffic data from permanent sensors*. 2021. URL: https://parisdata.opendatasoft.com/explore/dataset/comptages-routiers-permanents.

[21] Google. *Environmental Insights Explorer (EIE)*. 2022. URL: https://insights.sustainability.google.

[22] C. Gately, L.R. Hutyra, and I.S. Wing. *DARTE Annual On-road CO2 Emissions on a 1-km Grid, Conterminous USA V2, 1980-2017*. en. 2019. DOI: 10.3334/ORNLDAAC/1735. URL: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1735.

[23] Kaiming He et al. "Deep residual learning for image recognition". In: *CVPR*. 2016.

[24] Pesaresi M. and Politis P. *GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975-2030)*. European Commission, Joint Research Centre (JRC), 2022. DOI: 10.2905/D07D81B4-7680-4D28-B896-583745C27085. URL: http://data.europa.eu/89h/d07d81b4-7680-4d28-b896-583745c27085 (visited on 09/16/2022).

[25] Schiavina M., Freire S., and MacManus K. *GHS-POP R2022A - GHS population grid multitemporal (1975-2030)*. European Commission, Joint Research Centre (JRC), 2022. DOI: doi:10.2905/D6D86A90-4351-4508-99C1-CB074B022C4A. URL: http://data.europa.eu/89h/d6d86a90-4351-4508-99c1-cb074b022c4a (visited on 09/16/2022).

[26] US Federal Highway Administration. *State Motor-Vehicle Registrations*. 2018. URL: https://www.fhwa.dot.gov/policyinformation/statistics/2018/mv1.cfm.

[27] World Health Organization. *Global Status Report on Road Safety 2018*. Genève, Switzerland: World Health Organization, Jan. 2019.

[28] World Bank Group. *CURB: Climate Action for Urban Sustainability, Version 2.1*. 2019. URL: https://datacatalog.worldbank.org/search/dataset/0042029 (visited on 08/30/2022).

[29] United States Environmental Protection Agency (EPA). *GHG Emissions Factors Hub*. 2022. URL: https://www.epa.gov/system/files/documents/2022-%2004/ghg_emission_factors_hub.pdf (visited on 08/30/2022).

[30] Florczyk A. et al. *GHS Urban Centre Database 2015, multitemporal and multidimensional attributes, R2019A*. 2019. URL: https://data.jrc.ec.europa.eu/dataset/53473144-b88c-44bc-b4a3-4583ed1f547e.

[31] Global Administrative Areas. *GADM database of Global Administrative Areas, version 4.0*. 2022. URL: https://www.gadm.org (visited on 08/30/2022).

[32] Robert W Marx. "The TIGER system: automating the geographic structure of the United States census". In: *Government Publications Review* (1986).

## A OpenStreetMap Road Type Mapping

Table 3: Mapping of the three road types used in our emissions calclution to their corresponding OpenStreetMap [11] tags.

| Road Class | OpenStreetMap Tags |
|---|---|
| Highway | motorway, motorway_link, trunk, trunk_link |
| Arterial | primary, primary_link, secondary, secondary_link |
| Local | tertiary, tertiary_link, residential, living_street, unclassified |

## B Machine Learning Model Descriptions

**S2+OSM** Our baseline architecture consists of an MANet semantic segmentation model [14] with an EfficientNet-b3 backbone [15], trained using a per-pixel mean squared error (MSE) loss. Models are trained until convergence, measured using the validation loss. We select the model with the lowest validation loss for evaluation.

**S2+OSM Ensemble** The Sentinel-2 and OSM ensemble uses three different backbone models: EfficientNet-b3 [15], ResNet-34, and ResNet-101 [23]. Models were trained using the same six channel RGB + OSM input images as used in the S2+OSM model. Initial training showed improved performance from averaging the logits of each of these networks instead of the predictions, and all models were trained using MSE loss.

**Planet+OSM** The Planet and OSM model was trained using the same architecture, backbone, and stopping criteria as the S2+OSM model. The input was a six channel image consisting of RGB PlanetScope imagery and rasterized OSM road data. The Planet+OSM model was also used to explore the importance of the road versus off-road pixels. This was accomplished by separating the loss terms using the OSM data. The loss terms for the road pixels or off-road pixel term were multiplied by a factor of three. The results in Table 4 show that RMSE is decreased when the road pixel loss term is weighted higher, but an increase in MPE and MAPE suggesting off-road pixels are being predicted incorrectly.

Table 4: Evaluation of Planet+OSM models trained with various loss functions.

| Loss Modification | RMSE | MAPE | MPE | Pearson's $\rho$ |
|---|---|---|---|---|
| Standard MSE | 3329.9 | 159.9% | **41.01%** | **0.60** |
| Weight Off Road Loss | 3494.4 | 169.7% | 43.93% | 0.58 |
| Weight Road Loss | **3243.3** | 175.1% | 62.12% | **0.60** |

**GNN OSM** The GNN OSM models use a GATv2 network architecture with 14 layers, 2 attention heads, and 64 hidden channels. The OSM road network is initially represented as a multi-digraph, with each edge representing a directional road segment and nodes representing intersections. Road length, number of lanes, road type (highway, arterial or local) and link road indication (used for road segments such as sliproads and ramps) are used as features for the road segments. The road network is then converted to a line graph, inverting the graph's nodes and edges. Two additional features are computed for the edges connecting different road segments, representing the dot product between the segments' unit vectors at the point of intersection and the dot product of the segments' unit vectors for the overall direction of the segment.

As AADT values can span many orders of magnitude, the GNN model is trained to predict log AADT values. The loss function used to train the GNN model has two parts; the first is an L1 loss on estimated log AADT values when AADT ground truth is available. However, AADT ground truth annotations are fairly sparse (typically representing single digit percentages of the road network), so an additional consistency loss is added. The consistency loss is averaged over all the road intersections, and is calculated as a function of the total AADT values into and out of each intersection:

$$L_c = \frac{|\sum AADT_{in} - \sum AADT_{out}|}{\frac{1}{2}(\sum AADT_{in} + \sum AADT_{out})} \tag{1}$$

**GNN OSM+GHSL** The GNN OSM+GHSL model uses additional features derived from the GHSL BUILT-S [24] and GHSL POP [25] datasets. The GHSL BUILT-S dataset is a global raster dataset that provides a measure of how much of the Earth surface is built-up, measured in square meters per grid cell. The GHSL POP dataset is a global raster dataset that provides population density estimates. The GHSL POP dataset is converted to an estimated vehicle density by multiplying the population density by vehicles per-capita statistics for US states [26] or countries [27]. The rasters are sampled every 100m along each road segment and averaged to provide two additional features for each road segment.

**GNN OSM+CNN** The GNN OSM+CNN model uses additional features extracted from the trained S2+OSM model. The S2+OSM model is run over the Sentinel-2 imagery for each city, and the 16 features from the penultimate layer of the model are sampled at the center pixel of each road segment.

**GNN Ensemble** The GNN ensemble is a simple average of AADT estimates from five different GNN OSM models of varying model depths, ranging from 14 to 20 layers.

## C Emissions Calculation Details

Predicted road activity data (AADT) is used as the basis for our emissions calculation. We derive a city-specific emissions factor for each supported road type (highway, arterial, and local), based on several related factors: vehicle fleet mix, fuel types, fuel efficiencies, and GHG emissions factors

(EFs). When combined appropriately, these values can be converted to an emissions factor per road type in units of tonnes $CO_2$ per vehicle kilometer traveled (VKT). Then, AADT predictions for each road segment are multiplied by the length of that road segment to derive the estimated VKT for an entire year on that road. Multiplying the VKT activity by the EF for this road type provides the estimated emissions for this road. This process is repeated for all road segments under consideration within a city's bounds, as depicted in Figure 3.

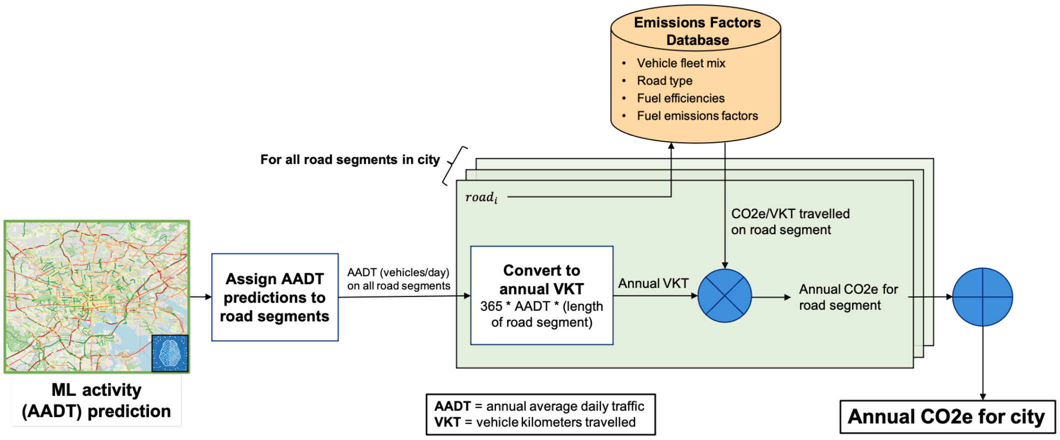
Emissions Factors Database
Vehicle fleet mix
Road type
Fuel efficiencies
Fuel emissions factors
For all road segments in city
$road_i$
CO2e/VKT travelled on road segment
ML activity (AADT) prediction
Assign AADT predictions to road segments
AADT (vehicles/day) on all road segments
Convert to annual VKT
365 * AADT * (length of road segment)
Annual VKT
Annual CO2e for road segment
AADT = annual average daily traffic
VKT = vehicle kilometers travelled
Annual CO2e for city

Figure 3: Emissions calculation overview, from ML-predicted AADT to emissions estimate for an entire city.

Data collection for each of these EF-related variables across 500 cities is a significant undertaking. The initial version of estimated emissions factors focused on collecting data at the country level for the 86 countries in which the top 500 cities are located. Sources for each type of data required for the emissions factor calculation are shown in Table 5.

Table 5: Data sources for variables used in emissions calculation.

| Data Type | Source(s) |
|---|---|
| Road Segment Type | OpenStreetMap [11] |
| Vehicle Fleet Mix | Various, available upon request |
| Fuel Type | CURB [28] |
| Fuel Efficiencies | CURB [28] |
| GHG Emissions Factors | US EPA GHG Emissions Factors Hub [29] |

# D Top 500 Cities Selection

To prioritize the set of 500 global cities, we utilized the European Union Joint Research Center Global Human Settlement Layer Urban Centers Database (GHSL-UCDB) dataset [30] for a globally consistent representation of city extent. This database contains the geographic bounds and other metadata for approximately 13,000 cities worldwide, and utilizes a definition of city/urban center based on population density and built up area. Specifically, an urban center was defined as "the spatially-generalized high-density clusters of contiguous grid cells of 1 km$^2$ with a density of at least 1,500 inhabitants per km$^2$ of land surface or at least 50% built-up surface share per km$^2$ of land surface, and a minimum population of 50,000" [30]. Due to this definition, city geometries in UCDB often have significantly different shapes and sizes as compared to official administrative bounds, e.g., from OSM [11] or Global Administrative Areas (GADM) [31]. We note that these differences are likely a main cause of discrepancies between our emissions estimates and other inventories.

UCDB spatially combines urban centers with a variety of metadata related to geography, socio-economic, environment, disaster risk, and sustainable development goals. This metadata includes EDGAR V5.0 [6] emissions estimates within urban center bounds for 1975, 1990, 2000, and 2015. We used the 2015 transport sector total CO2 emissions from non-short-cycle organic fuels (fossil fuels, `CO2_excl_short-cycle_org_C` in EDGAR) to sort and select the largest 500 cities for this work. The distribution of the selected cities across continents is shown below in Table 6.

Table 6: Regional distribution of the 500 global cities selected for emissions estimation.

| Region | Proportion of Top 500 Cities |
|---|---|
| Asia | 42.6% |
| Europe | 18.8% |
| North America | 17.8% |
| Latin America and the Caribbean | 11.2% |
| Africa | 8.2% |
| Oceania | 1.4% |

# E Metric Definitions

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - GT_i)^2} \tag{2}$$

$$\text{Mean Error} = \frac{1}{n}\sum_{i=1}^{n}(P_i - GT_i) \tag{3}$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \frac{100}{n}\sum_{i=1}^{n}\frac{|P_i - GT_i|}{GT_i} \tag{4}$$

$$\text{Mean Percentage Error (MPE)} = \frac{100}{n}\sum_{i=1}^{n}\frac{P_i - GT_i}{GT_i} \tag{5}$$

These metrics are used to analyze both predicted AADT and city-level emissions. In each equation, $n$ represents the number of roads/cities under consideration, $i$ represents the road/city index, $P$ represents the predicted AADT/predicted emissions from our model, and $GT$ represents the ground-truth AADT/emissions value.

## F List of U.K. 26 Cities

Table 7: List of U.K. 26 cities used for AADT evaluation.

London
Manchester
Birmingham
Leeds
Liverpool
Glasgow
Newcastle upon Tyne
Sheffield
Nottingham
Bristol
Portsmouth
Middlesbrough
Coventry
Edinburgh
Huddersfield
Southampton
Reading
Southend-on-Sea
Warrington
Runcorn
Wishaw
Blackburn
Atherton
Crawley
Slough
Coatbridge

## G U.S. Emissions Validation

Google Environmental Insights Explorer (EIE) [21] leverages trip data in combination with emissions factors data to provide emissions estimates for multiple modes of transportation in 42,000+ cities worldwide. We utilize the publicly available 2018 EIE data in the US for our comparison. DARTE [22] uses reported vehicular traffic data combined with Census TIGER [32] road network information to estimate regional on-road emissions and disaggregate them among mapped road networks. We compare our estimates to both DARTE 2015 and 2017 data. Vulcan [5] is a national-scale, multi-sectoral, hourly inventory from 2010-2015 with a resolution of 1 km$^2$. Vulcan transportation emissions are based on EPA county-level on-road emissions estimates, further downscaled using data from the Federal Highway Administration. We select Vulcan data from 2015 for comparison.

Due to the fact that our ground truth AADT data and satellite imagery for this set of cities is from 2017, data from the other emissions inventories were selected from years as close to 2017 as possible. We use the geographic bounds available in the EIE data to retrieve satellite imagery and road network data within each city's bounds. After predicting AADT with our models and associating AADT with each road segment, road geometries are cropped to the city bounds to create an appropriate estimate of VKT and emissions for each road. Corresponding emissions estimates from the DARTE and Vulcan raster products are also selected using each city's EIE bounds.

Table 8: Emissions validation metrics for US cities. MAE and Mean Error are in units of tonnes $CO_2$, and $\rho$ is Pearson's $\rho$.

| | **CNN** | | | | **GNN** | | | |
|---|---|---|---|---|---|---|---|---|
| Emissions Dataset | RMSE | Mean Error | MAPE | $\rho$ | RMSE | Mean Error | MAPE | $\rho$ |
| EIE_v1_2018 | 544,225 | 407,997 | 77.3% | 0.94 | 3,706,827 | 3,706,827 | 321.5% | 0.95 |
| EIE_v2_2018 | 1,180,437 | -1,153,065 | 36.1% | 0.94 | 2,223,303 | 2,145,764 | 71.3% | 0.96 |
| DARTE_2015 | 2,606,389 | -2,606,389 | 53.6% | 0.98 | 708,514 | 692,440 | 19.8% | 0.99 |
| DARTE_2017 | 3,505,472 | -3,505,472 | 59.5% | 0.98 | 875,254 | -206,642 | 17.2% | 0.99 |
| VULCAN_lo_2015 | 912,134 | 912,134 | 124.8% | 0.98 | 4,210,964 | 4,210,964 | 473.2% | 0.99 |
| VULCAN_mn_2015 | 761,213 | 759,672 | 93.3% | 0.98 | 4,058,502 | 4,058,502 | 391.8% | 0.99 |
| VULCAN_hi_2015 | 617,740 | 607,210 | 71% | 0.98 | 3,906,040 | 3,906,040 | 330.7% | 0.99 |

Several variants of each third party inventory are examined. Google EIE data categorizes trips into three categories: in-boundary, inbound, and outbound. Trips are categorized according to their start and end locations, with in-boundary containing trips that both start and end within city bounds, inbound starting outside and ending inside city bounds, and outbound starting inside and ending outside city bounds. We compare against just in-boundary emissions (EIE_v1_2018), and in-boundary plus 50% inbound and 50% outbound emissions (EIE_v2_2018). For DARTE, we compare against emissions estimates for both 2015 (DARTE_2015) and 2017 (DARTE_2017). Vulcan contains three emissions estimates: the lower 95% confidence interval (VULCAN_lo_2015), mean estimate (VULCAN_mn_2015), and the upper 95% confidence interval (VULCAN_hi_2015).

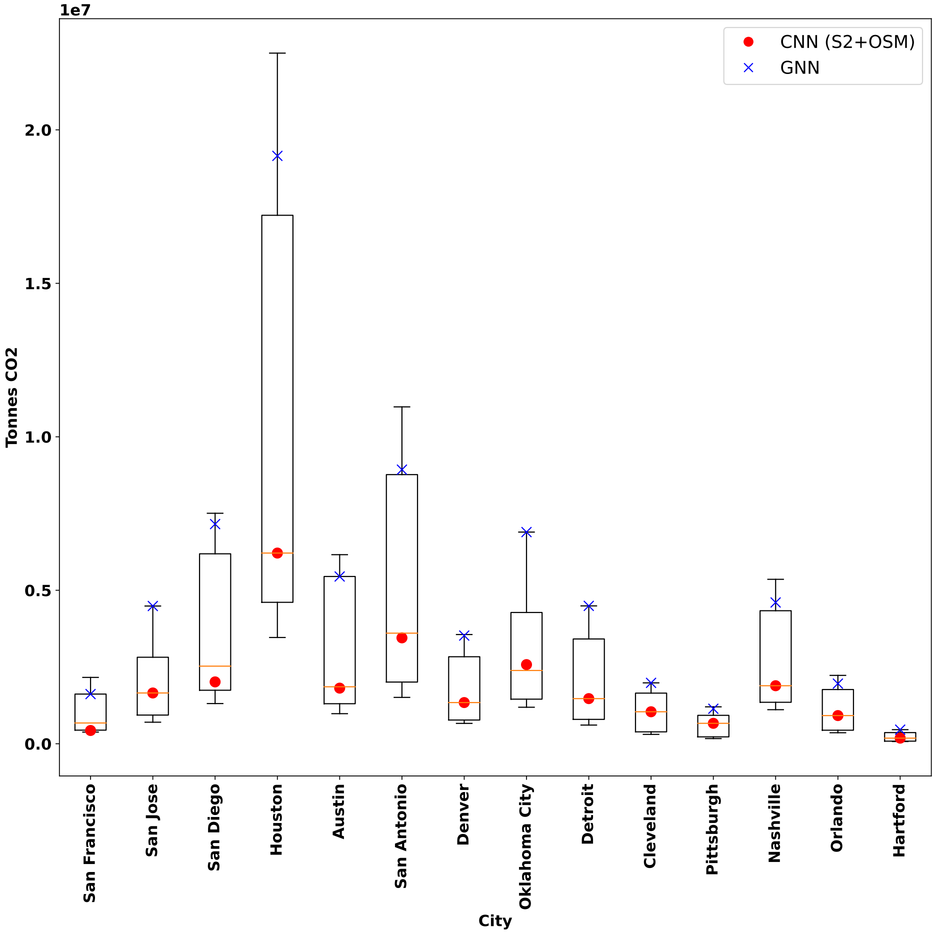

Figure 4: Distribution of emissions estimates for each US city. Emissions estimates based on our S2+OSM CNN are marked with red dots, and estimates based on our GNN OSM model outputs are marked with blue X's.

## H Global Emissions Validation

Initial validation was performed for our global emissions estimates, where we compare against both EDGAR [6] and Carbon Monitor city-level data for 2018-2020 [7]. EDGAR 2015 data is retrieved from the Global Human Settlement Layer - Urban Centres Database (GHSL-UCDB) dataset [30] from which we have selected our set of 500 global cities, and we acknowledge that more recent EDGAR data from 2018 should be used in future validation experiments. Carbon Monitor is a recent emissions inventory that utilizes a variety of activity data sources to estimate emissions in multiple sectors on a daily basis. In addition to country level data, Carbon Monitor has released near real-time emissions estimates for 52 cities globally. This city-level data is used in our analysis, for 50 total cities that overlap the global set of 500 cities for which we have produced emissions estimates.

Validation metrics for both dataset comparisons shown in Table 9. The resulting comparison for all 500 cities against EDGAR can be seen in Figure 5. While the Pearson $\rho$ value of 0.74 indicates decent correlation, the wide variance of the differences is noteworthy and warrants further investigation. The sharp “wall" on the left portion of the plot is caused by the fact that our 500 cities were selected based on thresholded EDGAR 2015 estimates.

Table 9: Global emissions validation metrics for our estimates compared with EDGAR [6] and Carbon Monitor [7] data. MAE and Mean Error are in units of tonnes $CO_2$.

| Emissions Dataset | # of Cities | MAE | MAPE | Mean Error | MPE | Pearson's $\rho$ |
|---|---|---|---|---|---|---|
| EDGAR 2015 | 500 | 1,158,740 | 68.80% | 248,624 | 23.60% | 0.74 |
| Carbon Monitor 2019 | 50 | 2,857,690 | 72.40% | -844,634 | 44.40% | 0.87 |
| Carbon Monitor 2020 | 50 | 2,634,598 | 83.20% | -317,283 | 55.70% | 0.86 |
| Carbon Monitor 2021 | 50 | 2,795,294 | 73.50% | -781,053 | 42.40% | 0.87 |

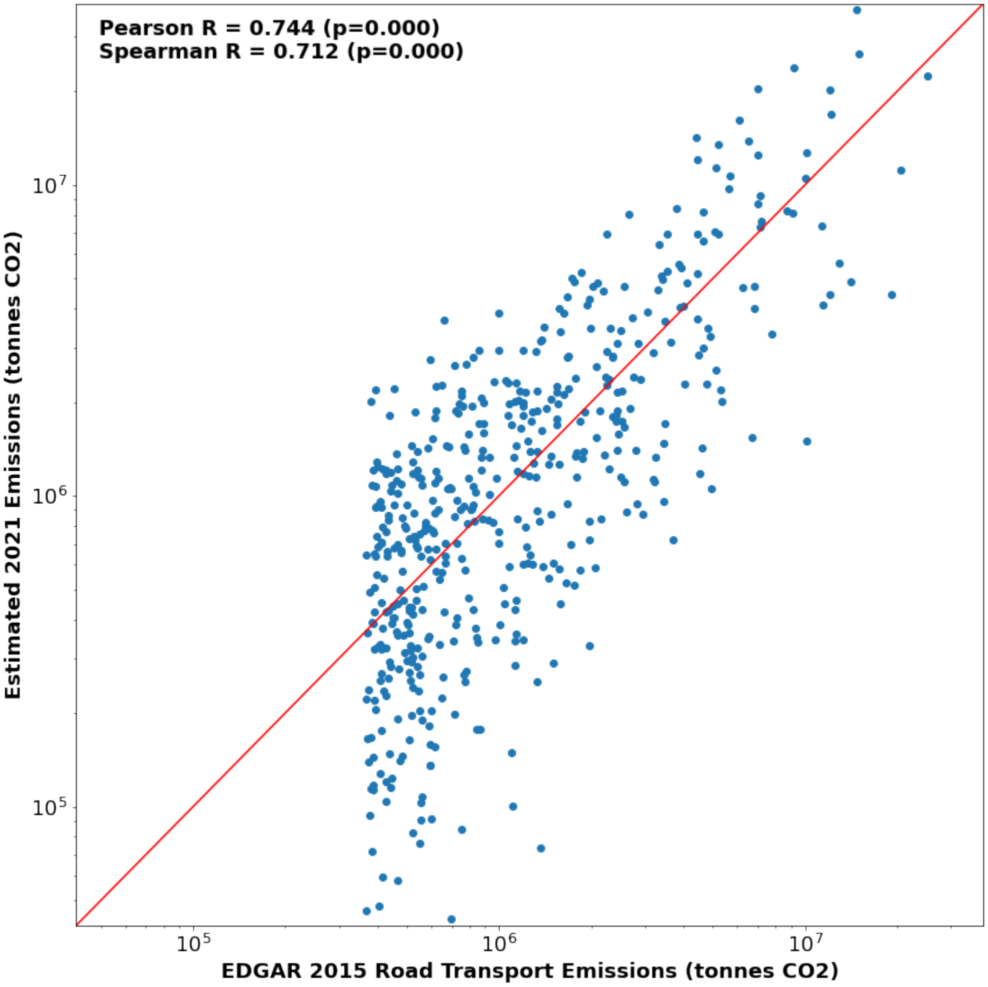

Figure 5: Our emissions estimates for 500 global cities compared with EDGAR 2015 data. Note that axes are in log scale.

The results of the comparison with the 50 overlapping Carbon Monitor cities for 2021 is shown in Figure 6. There is generally good alignment between the two sets of emissions, with some larger differences in France (Nice, Lyon, Marseille), South America (Bogota, São Paulo), Russia (Saint Petersburg, Moscow), and India (Mumbai, Delhi). We also note the larger percentage errors for 2020 in Table 9 as compared to 2019 and 2021, likely due to COVID-19 lockdown effects.

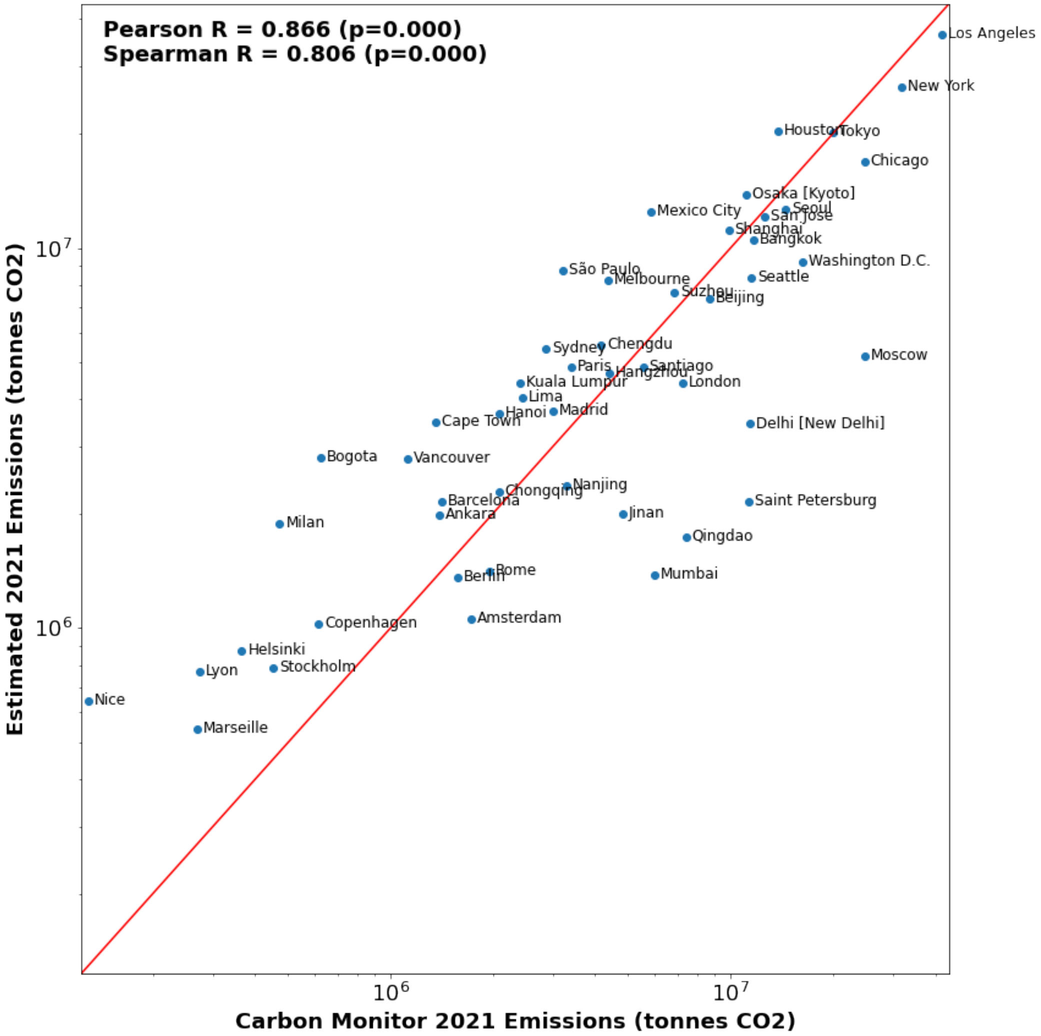

Figure 6: Our emissions estimates for 50 global cities compared with Carbon Monitor 2021 data. Note that axes are in log scale.