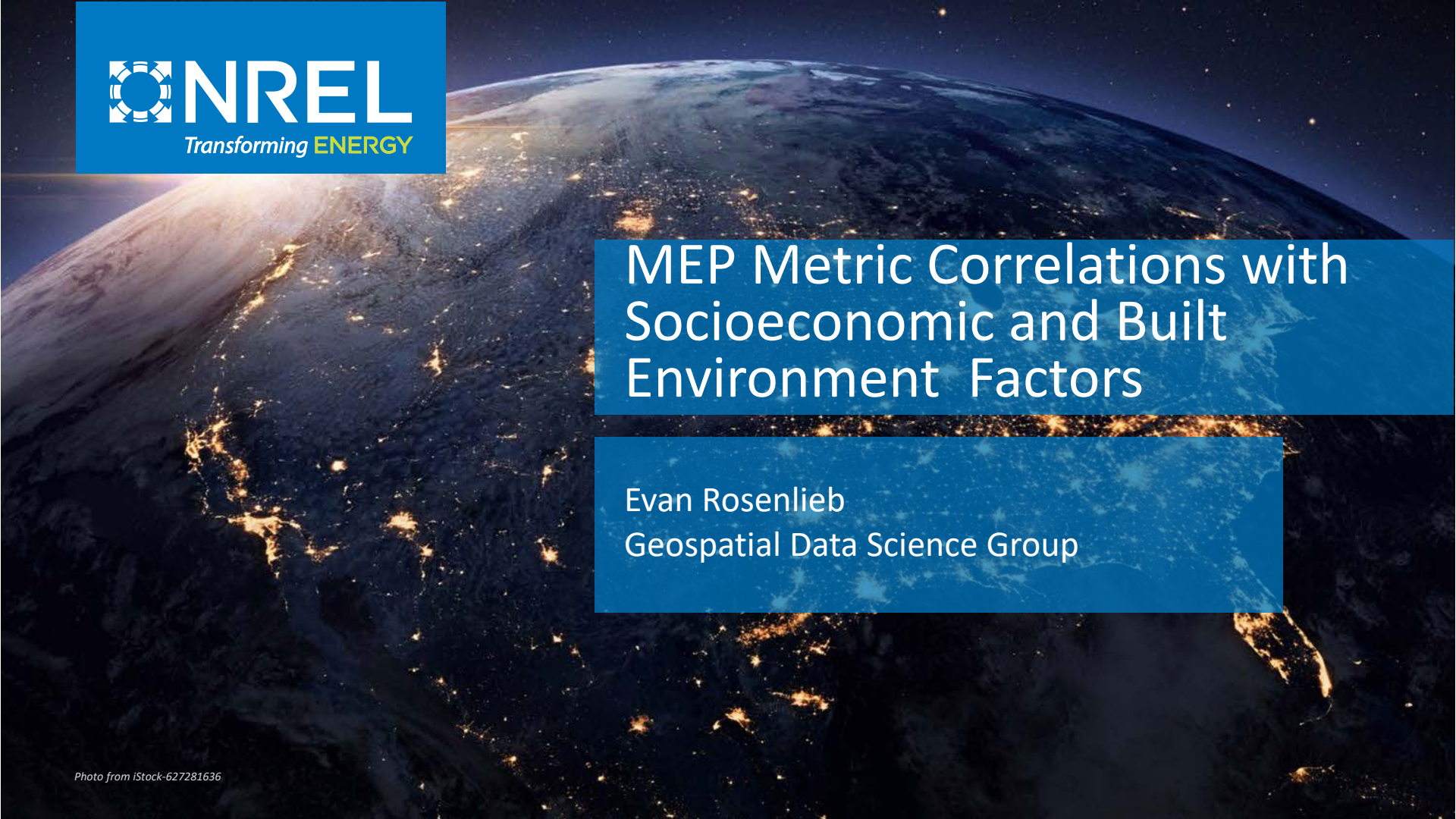


# MEP Metric Correlations with Socioeconomic and Built Environment Factors

Evan Rosenlieb
Geospatial Data Science Group

*Photo from iStock-627281636*

# Contents

# Introduction and Motivation

# Analysis Goals

The analysis presented examines potential relationships between the Mobility Energy Productivity (MEP) metric generated by NREL and other social, economic, demographic, and built environment variables with potentially interesting relationships with mobility in urban areas.

# The Mobility Energy Productivity Metric (MEP)

The MEP metric is a novel method of quantifying “the ability of an area's transportation system to connect individuals to goods, services, employment opportunities, and other activities while accounting for time, cost, and energy of modes that provide mobility in that area. ” It was developed by the TRAnsportation Modeling and Metrics (TRAMM) team at NREL.


**A Novel and Practical Method to Quantify the Quality of Mobility: The Mobility Energy Productivity Metric**

**Preprint**

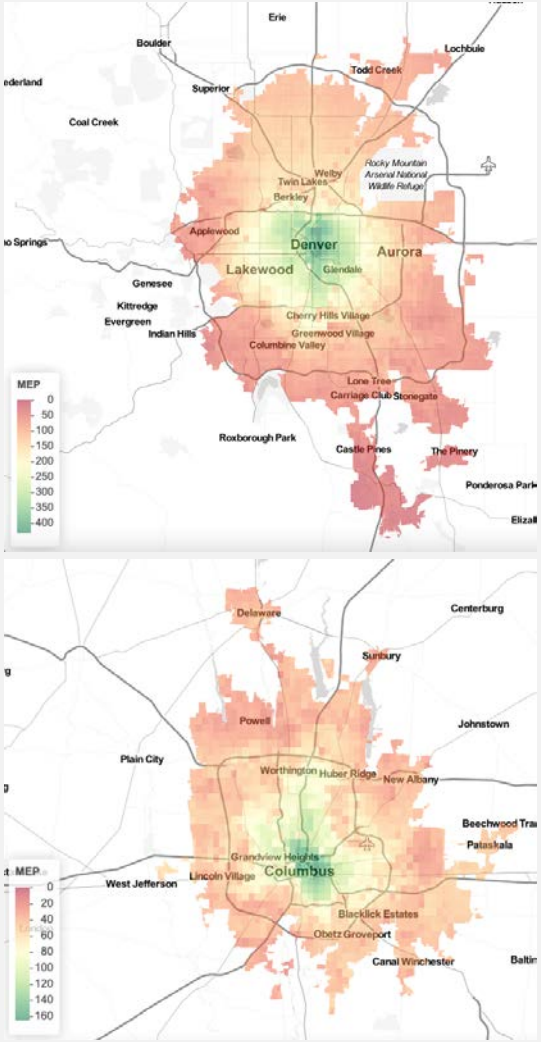
Yi Hou, Venu Garikapati, Ambarish Nag, Stanley Young, and Tom Grushka

*National Renewable Energy Laboratory*

*Presented at Transportation Research Board (TRB) 98th Annual Meeting*
*Washington, DC*
*January 13–17, 2019*

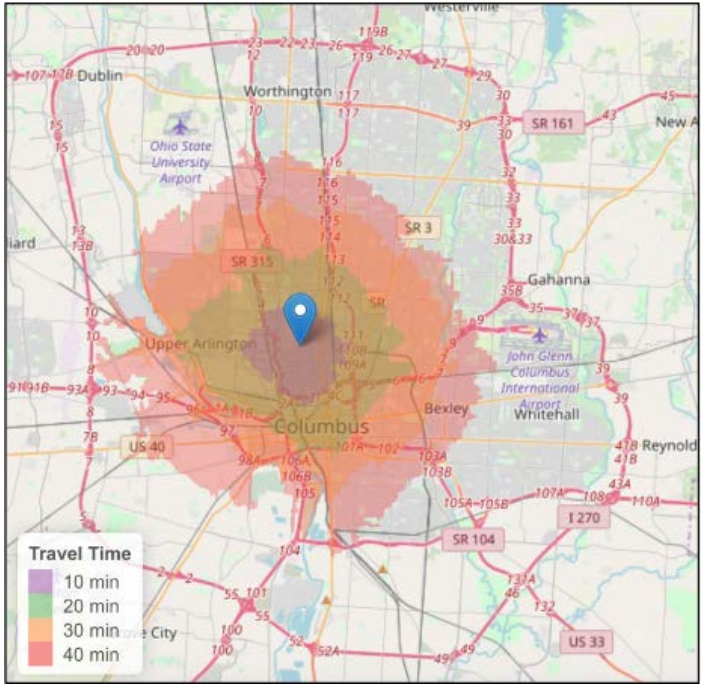
# The MEP project calculates this metric at sq.km grid resolution for US metropolitan areas.

Shown to the right: Examples of the MEP score for the Denver CO and Columbus OH metro areas are shown. The variation both within and between metro areas can be observed. Note: the MEP maps shown here are more recent versions than those used in the analysis. The version used in the analysis was calculated at the Census Bureau Metropolitan area boundary, whereas values shown are calculated at the Census Bureau Urbanized Area Boundary



# MEP Spatial Methods

- The MEP score works by quantifying each locations access to jobs and amenities via different modes of transportation.
- The more jobs and amenities (restaurants, grocery stores, educational facilities, recreational opportunities, hospitals) that can be easily reached by faster and lower energy as well as less cost intensive modes of transportation the better the score.
- The base modes of transportation used are walk, bike, transit, and car; and for each mode of transportation isochrones are computed up to 40 minutes of travel time with 10-minute increments.




Shown: Example 10 minute isochrones for the driving mode of transportation shown in Columbus, OH.

# MEP Spatial Methods

- The primary inputs to the MEP metric are transportation networks, data on energy and cost of transportation modes, and locations of jobs and amenities.
- While many ancillary datasets are used to properly weigh and apply these variables; the metric does not explicitly include as an input common sociodemographic or built environment metrics that are often used when describing urban areas, such as average income, racial and ethnic makeup, average commute time, or vehicle miles traveled.*

*While true of the MEP at this time of this analysis, version of the MEP have since been developed that incorporate some of these measures.

**In this project the relationship between the MEP score and socioeconomic and urban form variables, including those previously calculated for the C-LEAP project, was determined as a potential analysis of interest.**

- The fact that sociodemographic and built environment metrics are not inputs to the MEP model allows for a valid analysis of how they correlate with the metric. There are two potential reasons this may be desirable:
  - Knowledge of correlations with mobility productivity have equity implications: policy makers have an interest in providing equal access to high mobility productivity.
  - If other, more easily obtainable and computable factors can predict the MEP metric with high accuracy, they may be useful as a proxy to the full MEP score.
- By leveraging data created for the MEP project as well as high spatial resolution data created for the C-LEAP project, the impact of both datasets was extended through this analysis.

# Analysis Goals

An exploratory analysis was performed with two primary analysis goals:

- To understand which socioeconomic and built environment factors that appear significantly correlated with the MEP metric to identify, for instance, which sociodemographic groups have less access to energy efficient mobility. For this analysis a multiple regression analysis framework is used.
- To understand if factors can be identified that in total can accurately predict the MEP metric in areas where it has not yet been calculated. For this analysis a predictive random forest model is used.

Variables of interest are identified that are, whenever possible, available from a public data source with national coverage, such that the analysis could potentially be repeated anywhere in the US.

# Datasets Used

# Variables – Main Sources

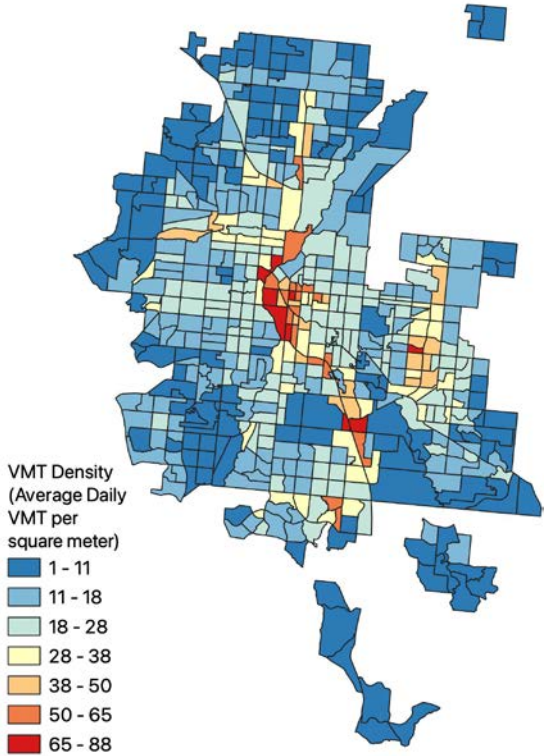
Relevant variables that met the criteria of being publicly available mostly belonged to a few broad categories:

- Demographic and Socioeconomic data obtained through the Census Bureau's American Communities Survey (ACS).
- Data on commute patterns from the Bureau of Transportation Statistic's Local Area Transportation Characteristics for Households (LATCH) survey.
- Data on transportation network density scraped from the open-source OpenStreetMap (OSM) project.

# Variables -- VMT

While most data used was available from these sources, a couple of other less accessible variables were identified as being important enough to explore:

- Total traffic density, as expressed by Vehicle Miles Traveled (VMT) per area is an important variable studied in urban studies and seems highly likely to be correlated with the MEP metric.
- VMT density is not publicly available at the fine-grained resolution MEP is calculated at.
- However, leveraging work previously done at NREL to estimate VMT by municipality for the C-LEAP project, such VMT density values were able to be imputed for certain areas using data from the Federal Highway Administrations Highway Performance Monitoring System (FHWA HPMS)



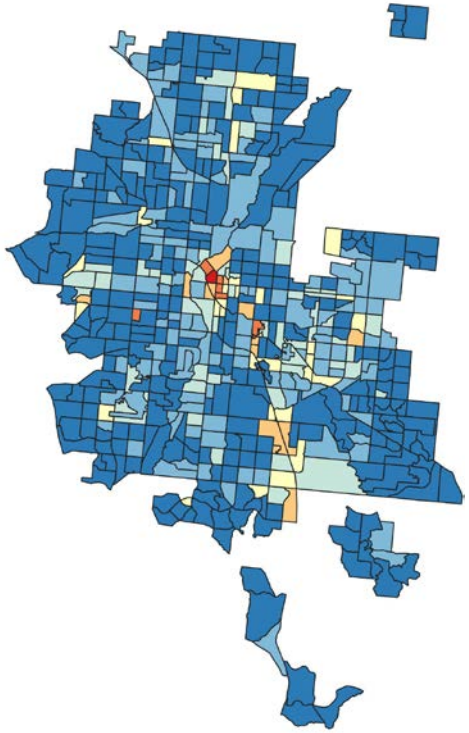

VMT density extracted to census tracks in Denver Metro Region

# Variables -- Parking

- Additionally, the density of parking spaces is a frequently studied variable in urban studies, as density of parking spaces is a measurement of physical infrastructure dedicated to vehicle commuting.

- Unfortunately, for this variable no comprehensive, high quality publicly available data source was able to be identified. Despite this, its high salience prompted the use of a proprietary third-party land use dataset for use in this analysis.

- While no nation wide publicly available dataset is available, it is likely that local policy makers have some data available to them for their location, meaning that potentially such data could still be used for individualized analyses.



Parking density extracted to census tracks in Denver Metro Region

# All Variables Used

| Variable | Source | Source Description |
|---|---|---|
| MEP Index | NREL | Calculated for most large metro areas |
| Roadway Vehicle Miles Traveled Density | NREL using HPMS data | Publicly available for all of US but data quality highly variable by state |
| Average Weekday Household Vehicle Trips | LATCH | Publicly available for all of US |
| Average Weekday Household Vehicle Miles | LATCH | Publicly available for all of US |
| Average Weekday Household Person Trips | LATCH | Publicly available for all of US |
| Average Weekday Household Person Miles | LATCH | Publicly available for all of US |
| Percent Persons Not Nonlatino-White | ACS | Publicly available for all of US |
| Percent Persons with Associates Degree or Higher | ACS | Publicly available for all of US |
| Percent Persons with Disability | ACS | Publicly available for all of US |
| Percent Persons 65 or older | ACS | Publicly available for all of US |
| Percent Persons receiving SNAP assistance | ACS | Publicly available for all of US |
| Percent Households Renting | ACS | Publicly available for all of US |
| Percent Households with no Vehicle | ACS | Publicly available for all of US |
| Percent Households with Four or More Vehicles | ACS | Publicly available for all of US |
| Percent Household living below Poverty line | ACS | Publicly available for all of US |
| Percent Persons with Commute under 24 Minutes | ACS | Publicly available for all of US |
| Percent Persons with Commute over 45 Minutes | ACS | Publicly available for all of US |
| Road Density | OSM | Open-source global dataset |
| Traffic Signal Density | OSM | Open-source global dataset |
| Bus stop Density | OSM | Open-source global dataset |
| Train Stop Density | OSM | Open-source global dataset |
| Parking Space Density | Proprietary | Proprietary, local alternatives may be available for a particular urban area |

# Variable Summary Statistics

| Variable | Minimum | Median | Max |
|---|---|---|---|
| MEP Index | 11.2 | 109.4 | 426 |
| Roadway Vehicle Miles Traveled Density | 1.3 | 16.1 | 118.4 |
| Average Weekday Household Vehicle Trips | 2.3 | 5.2 | 7 |
| Average Weekday Household Vehicle Miles | 14.1 | 34.7 | 60.2 |
| Average Weekday Household Person Trips | 4.9 | 8.1 | 11.4 |
| Average Weekday Household Person Miles | 25.2 | 50.1 | 80.5 |
| Percent Persons Not Nonlatino-White | 1.8 | 33.5 | 100 |
| Percent Persons with Associates Degree or Higher | 4.6 | 42.4 | 90.9 |
| Percent Persons with Disability | 1.8 | 10.9 | 33.4 |
| Percent Persons 65 or older | .5 | 11.3 | 35.3 |
| Percent Persons receiving SNAP assistance | 0 | 9.3 | 65 |
| Percent Households Renting | .6 | 40.1 | 100 |
| Percent Households with no Vehicle | 0 | 5.6 | 44.5 |
| Percent Households with Four or More Vehicles | 0 | 4.6 | 20.2 |
| Percent Household living below Poverty line | .2 | 12.3 | 70.9 |
| Percent Persons with Commute under 24 Minutes | 26.9 | 56.7 | 88.9 |
| Percent Persons with Commute over 45 Minutes | 0 | 11.8 | 32 |
| Road Density | 1.25 | 10.8 | 19.5 |
| Traffic Signal Density | 0 | 2.5 | 49.5 |
| Bus stop Density | 0 | .9 | 27.1 |
| Train Stop Density | 0 | 0 | 4.2 |
| Parking Space Density (values shown here are scaled by 1000) | 0 | .7 | 11.7 |

Variables with “density” in name are counts per square meter

# Spatial Methods

# C-LEAP VMT Data

- To produce Transportation energy use data by city, C-LEAP used FHWA HPMS data to estimate VMT for all road segments supplied by HPMS. Detailed estimates are only provided for highways and major roads by HPMS.
- However, traffic of surrounding collectors and local roads can be estimated using traffic data of surrounding  major roads and calibrating to state totals.
- Although this estimated data aggregated up to the city level for C-LEAP, it can easily be aggregated up to different geometries, without any interpolation or loss of data quality, allowing for comparison with MEP and other variables at any arbitrary spatial scale.

# Missing HPMS Data

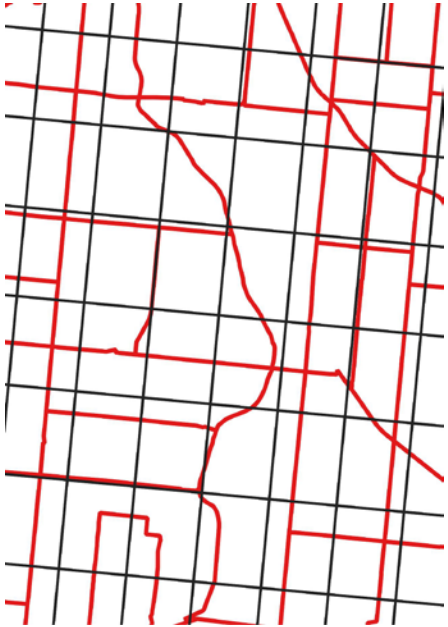
- To pilot this analysis, we started with the Denver Metro area as it was one of the first Metro areas the MEP score was computed for, and we had familiarity with the area.

- When we tried to expand past the Denver Metro area, we discovered that Denver was unique in that it had complete HPMS data. Most of the country does not have complete enough HPMS data to reliably calculate VMT at the tract level (spatial data is only provided for major arterials, highways and interstates).

- The next two largest Metro areas with complete VMT data were Birmingham AL and Indianapolis IN. Therefore, all analysis was completed on these three cities.

# Choosing a Unit for Spatial Analysis

- Broadly speaking, these datasets are available three units of spatial analysis:
  - The MEP score, calculated on its own equirectangular grid
  - All ACS and LATCH variables, which are provided at the unit of the census tract
  - VMT, OSM, and COSTAR datasets which are provided at discrete locations that can be aggregated up to different units of analysis.
- MEP grid cells are similar in size to small census tracts and much significantly smaller than large census tracts, this makes census tracts the lowest spatial resolution unit.
- Therefore, to harmonize all variables to one unit of analysis, variables not already associated with census tracts are spatially re-aggregated to the census tract level. In the case of the MEP score, the area weighted average of MEP grid cells that intersect with each census tract is used.

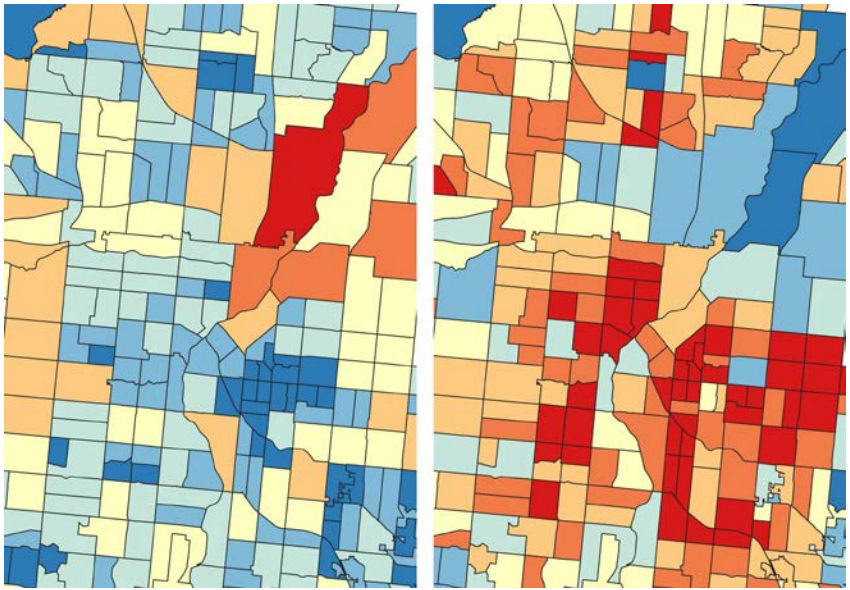


Comparison of MEP grid cells (in black) with census tracts (in red) in central Denver

# Totals vs Density

- Many of the metrics used in this analysis are measures of spatially dependent phenomenon. As such, they will potentially fall prey to the Modifiable Areal Unit Problem (MAUP).
- A large census tract will tend to have have more parking spaces than a small one simply because it is large enough to fit more.
- The size of the unit of analysis is an artifact of the census process and is not of interest to this analysis – therefore all variables related to aspects of the transportation infrastructure are normalized by tract area, so they are expressed as densities rather than totals.
- For a similar reason, the MEP by tract is calculated through a weighted average instead of a weighted sum.

Road Network Length Sum vs Density:



Sum

Density

# Combining City Data

- The MEP scores are very different from city to city – simply combining the data from cities lead to poor models because the only significant predictor was what city the point was from.
- The use of a city intercept variable was explored. However, this only controls for differences in the central tendency of the MEP score between cities, not the differences in variation.
- To sidestep this problem, the MEP scores were normalized for each city by taking the Z-score before combining. In this way, the models are seeing the extent to which the MEP score, *relative to its metro area,* can be predicted from the other variables.
- This normalization means that the resultant model cannot predict the absolute MEP score for a census tract that could be directly compared with other metro areas, however within-metro area comparisons are in most cases likely important for local policy makers than between-metro area comparisons.

# Regression Analysis

# Regression Analysis

- A multiple regression analysis is used to identify statistically significant relationships between the MEP score and other variables:
  - A regression analysis is performed with the metropolitan area MEP Z-score being regressed upon other variables of interest as predictors.
  - Model selection and diagnostic criteria are used to remove issues of multicollinearity and leverage and choose the most parsimonious models.
  - The significance and measures of fit of output models are examined, as well as the effect size of individual predictors.

# Variable Pre-Screening

- Correlation and scatterplot matrices were observed to spot variables requiring potential transformations and pairwise variables that are highly colinear.
- With the requisite variables expressed as densities in respect to area and the MEP scores expressed as Z-scores, no variables seem to present extreme skew that would require transformation.
- Some groups of variables exhibited extremely high multicollinearity with each other, namely:
  - % commute < 24 minutes and % commute > 45 minutes
  - % households 0 vehicle and % households 4 or more vehicles
  - Average weekday vehicle trips, vehicle miles, person trips, and person miles
- In these cases – the variable was chosen that had higher Pearson's R with MEP density to continue with analysis.
- Out of 881 census tracts in the study area, initial diagnostic criteria show a single census tract with very high potential leverage, so this tract (covering the 16$^{th}$ street mall in downtown Denver) is removed before model selection.

# Model Selection

- Because neither the number of variables nor the number of observations was exceedingly high, it was possible to perform an exhaustive model selection algorithm to compare all possible models that could be built with the selected variables.
- Using the AIC criteria selects a model with 12 out of 16 input variables, but results in a model with multiple nonsignificant predictors.
- Using the more aggressive BIC criteria results in a model with 10 variables, with the model and all predictors significant at the .01 alpha level.

# Initial Recommended Model

| Predictor | Coefficient | P Value |
|---|---|---|
| (Intercept) | 1.669 | 4.94E-04 |
| **Percent Persons with Bachelors Degree or Higher** | 0.014 | 3.56E-18 |
| **Percent Persons 65 or Older** | -0.020 | 3.67E-04 |
| **Percent Households Renting** | -0.008 | 2.10E-04 |
| **Percent Household living below Poverty line** | 0.018 | 3.64E-08 |
| **Percent Persons with Commute over 45 Minutes** | -0.017 | 1.39E-04 |
| **Roadway Vehicle Miles Traveled Density** | 0.008 | 1.69E-04 |
| **Average Weekday Household Vehicle Trips** | -0.505 | 1.63E-15 |
| **Parking Space Density** | -55.475 | 7.04E-03 |
| **Road Density** | 0.057 | 7.67E-11 |
| **Bus stop Density** | 0.063 | 1.07E-21 |

Multiple R-squared:  0.5716, p-value: < 2.2e-16

# Multicollinearity Diagnostics

- Due the number of variables related to factors of the transportation system and interrelated demographic factors respectively, it would not be surprising for there to be a high degree of multicollinearity among predictors.
- The ”Condition Index” multicollinearity diagnostic returns a maximum index of 75.8, quite a bit higher than the rule of thumb of 30 that shows a serious multicollinearity problem.
- Because of this, while the model may have predictive value, the coefficients might not be interpretable. In order to see if a model with higher interpretability is possible, the predictor with the highest multicollinearity with other variables, average weekday vehicle trips, is removed and model selection is begun again.
- A new model is returned, with 7 predictors, again with the model and all predictors significant at the .01 alpha level.

# Recommended Model

This model has a very similar R-squared, ~.55 vs ~.57 which is a very small decline considering it has 3 fewer predictors. The condition index returns nothing of concern.

| Predictor | Coefficient | P Value |
|---|---|---|
| (Intercept) | -1.803 | 7.83E-36 |
| **Percent Persons with Bachelors Degree or Higher** | 0.012 | 2.67E-14 |
| **Percent Households with no Vehicle** | 0.023 | 3.01E-06 |
| **Percent Household living below Poverty line** | 0.020 | 1.18E-09 |
| **Percent Persons with Commute over 45 Minutes** | -0.020 | 4.31E-06 |
| **Roadway Vehicle Miles Traveled Density** | 0.009 | 1.94E-05 |
| **Road Density** | 0.069 | 1.27E-15 |
| **Bus stop Density** | 0.064 | 1.41E-23 |

Multiple R-squared:  0.5465, p-value: < 2.2e-16

# Recommended Model Partial Correlations

- Unsurprisingly – the variables that most closely track the simple presence of transportation infrastructure explain most of the variance of the model. Tracts with higher road and bus stop density have the highest MEP scores, with bus stop density having the highest explanatory power.
- Socioeconomic variables of educational attainment and poverty level are the next most important variables.
- There do appear to be real associations between socioeconomic factors and areas mobility energy productivity, even after accounting for other built environment factors, showing the potential for the MEP score as a tool to examine and address equity concerns at the policy level.

| Predictor | Partial Correlation |
|---|---|
| **Percent Persons with Bachelors Degree or Higher** | 0.253 |
| **Percent Households with no Vehicle** | 0.157 |
| **Percent Household living below Poverty line** | 0.203 |
| **Percent Persons with Commute over 45 Minutes** | -0.154 |
| **Roadway Vehicle Miles Traveled Density** | 0.143 |
| **Road Density** | 0.265 |
| **Bus stop Density** | 0.328 |

# K-Fold Cross Validation

- Another way to test if the model is overfitted, as well as to get a good sense for its predictive value, is to perform a k-folds cross validation. If the model is over fitted, then there may be drastic reductions in explanatory power when predicting removed deleted observations.
- A 5-fold cross validation is conducted on the recommended model. In addition, the 10-predictor model is also validated, as a test of the highest possible predictive linear model.
- Both models perform well, with RMSE remaining stable between validations. While the 10-predictor model is stable and does have a lower RMSE than the 7-predictor model, ultimately it only results in predicting the MEP score ~.02 Z-scores better on average.

| Resample | RMSE - 7 Variable Model | RMSE - 10 Variable Model |
|---|---|---|
| Fold1 | 0.641 | 0.684 |
| Fold2 | 0.665 | 0.683 |
| Fold3 | 0.705 | 0.632 |
| Fold4 | 0.657 | 0.629 |
| Fold5 | 0.698 | 0.658 |
| Full Model | 0.67 | 0.652 |

# Predictive Model

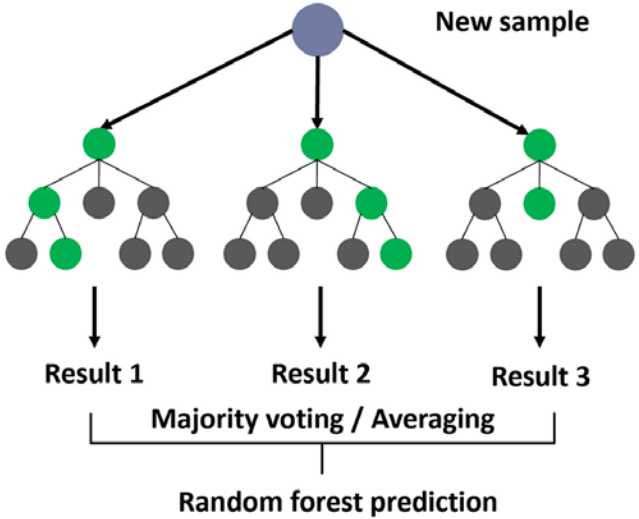
# Random Forest Model

- The results of the regression model show many identified variables are highly associated with the MEP score and together have some measure of predictive ability. However, if the goal is specifically prediction, the linear regression model can predict at most half of the variance of the MEP score at the tract level and are likely not predictive enough to act as a proxy for the actual MEP score for any reasonable application.

- Therefore, it is worthwhile to examine whether these variables can have any more predictive value when used in a different modeling framework. A random forest model is examined to see if the same variables can perform a better job at predicting the MEP score using a model geared specifically towards predictive power instead of interpretability.

# Random Forest Model

- The random forest model is a machine learning method that is designed to combine multiple decision trees trained on random samples of a dataset to predict an outcome variable with as high precision as possible without overfitting.
- Much like a regression model, a random forest model allows for an imputation of an outcome variable given the input variables that the model was trained on.
- The model relies on fewer fundamental assumptions about the data, such as linearity and lack of multicollinearity, meaning that it may be able to predict the outcome better for datasets that do not follow these assumptions.
- While it is possible to inspect the model to identify the most influential input variables or decision points, they do not provide the simple and intuitive coefficients between prediction and predictor as a linear regression.



# Random Forest Model

- To give this model as much information to work with, all identified variables are fed with no pre-screened variables.
- A random forest model called on all predictor variables produces a model with a RMSE of .623, lower than the .652 of the recommended linear model. Tuning of the model was not able to produce a model with a RMSE lower than .62.
- While R-squared values cannot be compared directly between models, for the sake of comparison a multiple regression model on this dataset with an RMSE of .62 would have an R-squared of .59 (vs .547 for the recommended regression model).
- While this represents a moderate increase in predictive power, and is higher than any regression model produced, it does not represent a large enough increase to likely make the outputs useful as a stand in for the MEP metric.

# Summary

- This work examined whether the Mobility Energy Productivity metric could be predicted at a census tract level buy a mixture of socioeconomic and built environment variables.
- Both linear regression and random forest models were used.
- After a model selection process, the recommended linear regression was significant overall with all seven predictor variables being significant as well. The model had moderate predictive power as shown by the R-squared of ~.55
- The random forest model was able to produce a better predictive model, however the difference in RMSE was marginal.

# Discussion

- The variables identified likely are not able to predict MEP scores with enough accuracy to be used as a stand in for any meaningful purpose. However, there is enough association with these publicly available variables it doesn't preclude the possibility that additional variables, perhaps of proprietary nature, could produce a model with enough predictive power to produce proxy MEP values with use in certain situations.

- Both built environment and socioeconomic variables are significantly correlated with the MEP score, even with controlling for each other, with socioeconomic variables explaining a fair amount of the variance.

- The fact that socioeconomic factors do seem to have some significant predictive power for the MEP score may have potential urban policy implications for the energy equity of transportation infrastructure.

# Thank You

**www.nrel.gov**

NREL/PR-6A20-86310



*Photo from iStock-627281636*

NREL
Transforming ENERGY