Using machine learning to understand causal relationships between urban form and travel CO_2 emissions across continents

Felix Wagner, Florian Nachtigall, Lukas Franken, Nikola Milojevic-Dupont, Nicolas Koch, Rafael H. M. Pereira, Jakob Runge, Marta C. Gonzalez, and Felix Creutzig

Content:

1. Background

Context, Gaps & Research Questions

1. Background, Gaps & Research Questions

Motivation: Built environment as a leverage point for change towards low carbon mobility

Background

- Urban transport responsible for **3 $GtCO_2$** per year (Creutzig et al. 2016)
- **Infrastructure** modifications **most relevant** for changing urban transport in comparison to personal or social factors (Javaid et al. 2020)
- currently unclear how to **translate IPCC's** national level **policies** into location-specific actions

Main Gaps

1. Causality of urban form effects
2. Scalability of recommendations
3. Location specific recommendations

1. Background, Gaps & Research Questions

Motivation: Built environment as a leverage point for change towards low carbon mobility

Background

- Urban transport responsible for **3 $GtCO_2$** per year (Creutzig et al. 2016)
- **Infrastructure** modifications **most relevant** for changing urban transport in comparison to personal or social factors (Javaid et al. 2020)
- currently unclear how to **translate IPCC's** national level **policies** into location-specific actions

Main Gaps

1. Causality of urban form effects
2. Scalability of recommendations
3. Location specific recommendations

Research Questions

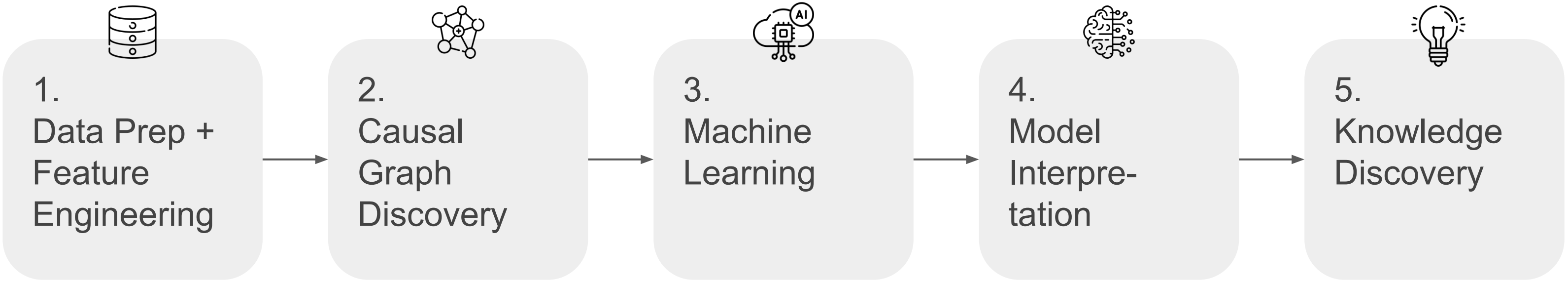
1. What is the causal relationship between the built environment and travel across cities?
2. What is the effect of individual built environment variables on trip emissions across cities?
3. What is the spatial heterogeneity of individual effects?

2. Methods

Pipeline Overview

2. Methods

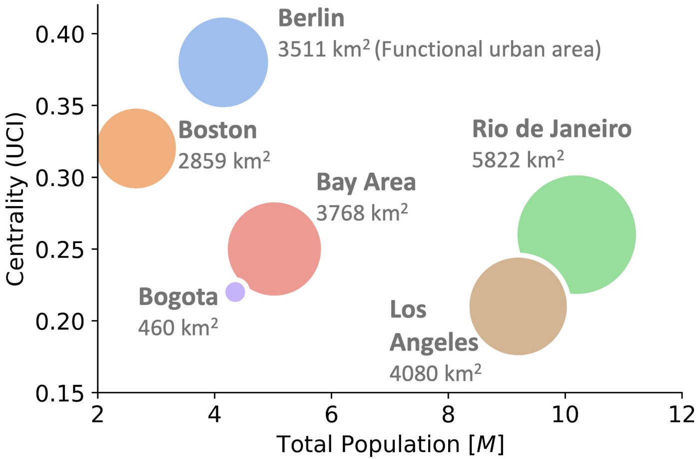
Framework to detect location specific effects



2. Methods

Framework to detect location specific effects

1. Data Prep + Feature Engineering → 2. Causal Graph Discovery → 3. Machine Learning → 4. Model Interpre-tation → 5. Knowledge Discovery



Target

Mean travel distance per TAZ

Features

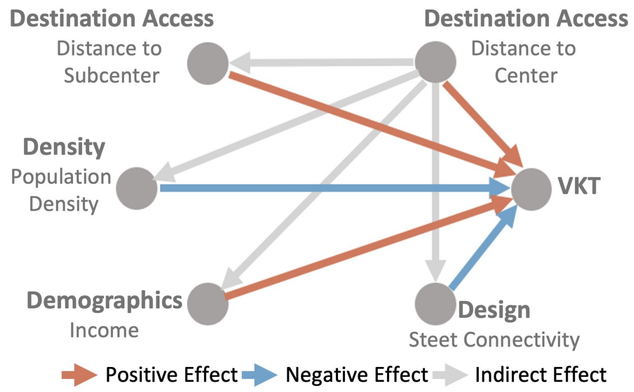
Destination access, Density, Design, Demographics

2. Methods

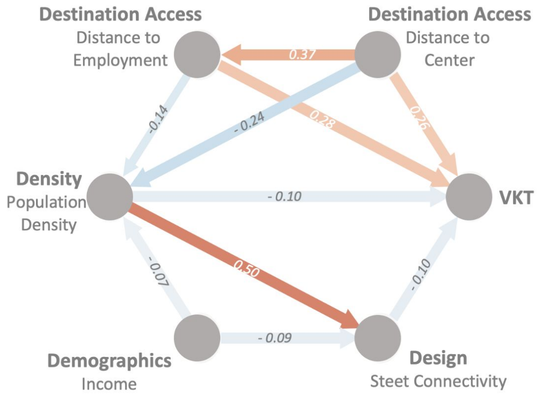
Framework to detect location specific effects

1. Data Prep + Feature Engineering → 2. Causal Graph Discovery → 3. Machine Learning → 4. Model Interpre-tation → 5. Knowledge Discovery

DAG based on literature



PC-based DAG



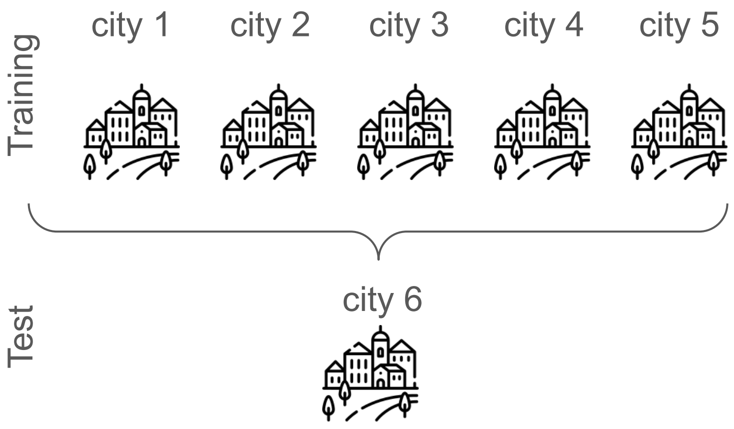
2. Methods

Framework to detect location specific effects

1. Data Prep + Feature Engineering → 2. Causal Graph Discovery → 3. Machine Learning → 4. Model Interpre-tation → 5. Knowledge Discovery

ML Model

- Gradient Boosting Decision Tree Regression Model (XGBOOST)
- only features with a direct causal effect on target (from DAG Discovery)
- 6-fold, city-wise cross validation
- Hyperparameter optimization per fold



2. Methods

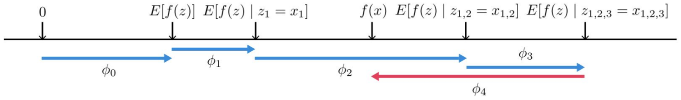
Framework to detect location specific effects

1. Data Prep + Feature Engineering → 2. Causal Graph Discovery → 3. Machine Learning → 4. Model Interpre-tation → 5. Knowledge Discovery

Causal Shapley Values (Heskes et al, 2022)

- Shapley Values: prediction score is distributed to a model's individual features
- **Benefit:** individual feature importance per sample (in our case: locations)
- **Difference:** Causal shapley values incorporate causal structure when distributing feature importance

(Marginal) Shapley Values:



2. Methods

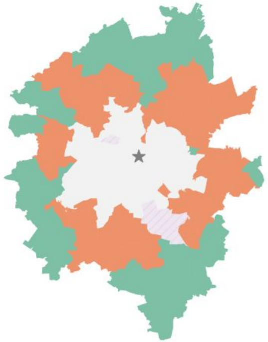
Framework to detect location specific effects

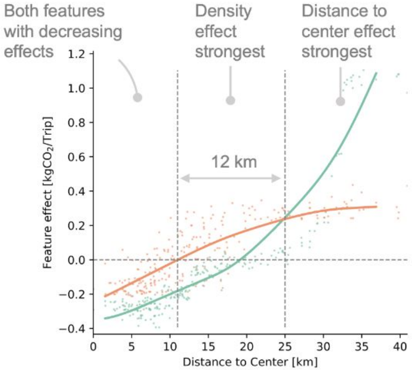
1. Data Prep + Feature Engineering → 2. Causal Graph Discovery → 3. Machine Learning → 4. Model Interpre-tation → 5. Knowledge Discovery

Where are urban form effects most significant?

-

A) Berlin





3. Causal Graph Discovery

Assumptions & Rationales

3. Causal Graph Discovery

Assumptions and rationales - how to communicate & test them?

Assumptions: Setup	**Rationales &** ***additional validation strategy***

3. Causal Graph Discovery

Assumptions and rationales - how to communicate & test them?

Assumptions: Setup	**Rationales & *additional validation strategy***
variables (x_1,x_2,...x_i, y)	based on previous urban form literature
causal discovery framework: PC algorithm - causal markov condition - faithfulness - causal sufficiency	- non-time series data, continuous variables w. linear & some nonlinear dependencies, different marginal distributions - ***test alternative framework (TBD)*** - ***leave-one-out analysis*** to assess influence of potentially missing nodes or variables outside the domain (Bönisch et al, 2023)
conditional independence (CI) test: Robust Partial Correlation CI test	- based on variable relationships - ***validation with alternative CI test (CMIknn)***

3. Causal Graph Discovery

Assumptions and rationales - how to communicate & test them?

Assumptions: Implementation	**Rationales &** ***additional validation strategy***
One graph across all cities	- general representation of urban form effects on travel - ***Comparison with one DAG per city*** - ***Remove city specific bias:*** balance sample & mean over several sampling rounds - ***Remove city specific confounding:*** normalise and standardise variables
Adding expert knowledge: - urban form cannot be caused by VKT - income cannot be caused by urban form (residential self selection) - distance to center cannot be caused by others	based on previous urban form literature & DAG literature
Comparison to DAG from literature	High dependence on our modelling decisions. Therefore, only analysis of differences.

Feedback? Questions?

References

1. Böhnisch et al (2023) European heatwave tracks: using causal discovery to detect recurring pathways in a single-regional climate model large ensemble. Environmental Research Letters

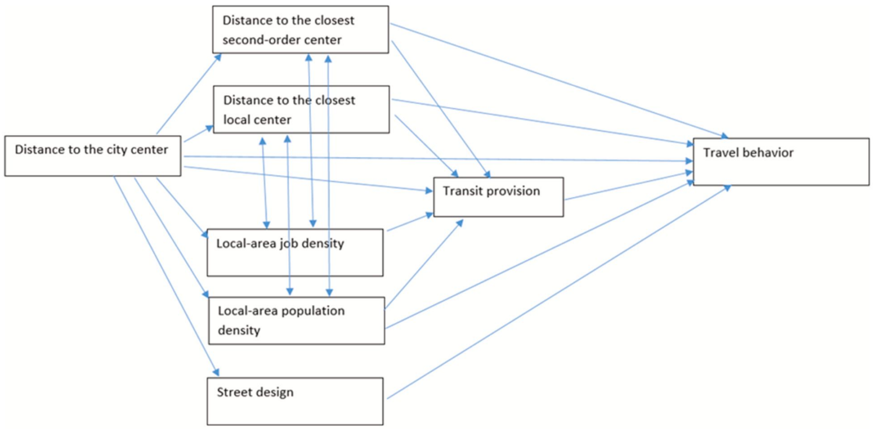
Appendix

1. Introduction
Gap 1: Causality

- **6D’s of compact development** for analysis of influence of built environment (BE) on car travel distance (VKT)
- **Urban form effects are not independent:** f.e. some D's on metropolitan level while others on neighborhood level
- only few studies reflected such dependencies
- **previous causality based studies have shortcomings:** cost intensive, hardly spatially representative

6D’s of compact development:

1. Destination Accessibility
2. Density
3. Diversity
4. Design
5. Distance to Transit
6. Demographics



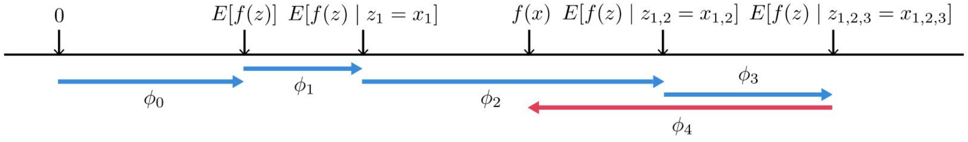
Naess et al. (2019)

2. Methods

2.3 Model Interpretation

- Interpretation via **Causal Shapley Values** (Heskes et al, 2022)
- Shapley Values (Lundberg and Lee (2017): prediction score is distributed to a model's individual features
- **Benefit:** individual feature importance is calculated per sample (in our case: locations)
- **Difference:** Causal shapley values incorporate causal structure (causal chain) when distributing feature importance

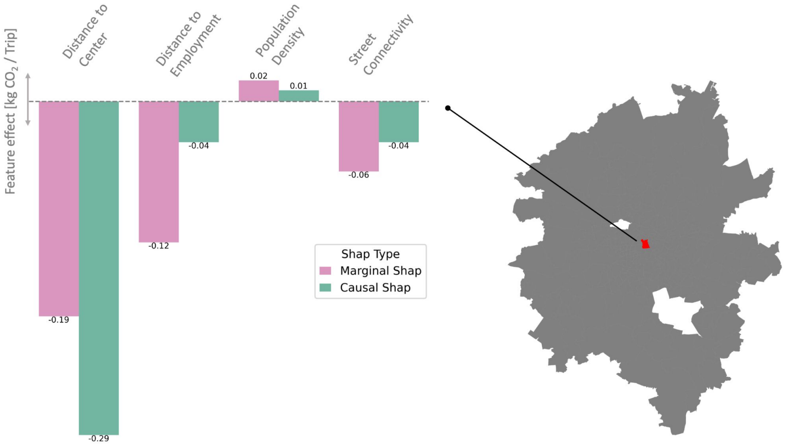
(Marginal) Shapley Values:



Lundberg and Lee (2017)

Causal Shapley Values:

Distance to Center -> Distance to employment -> Population Density -> Street Connectivity



3. Results

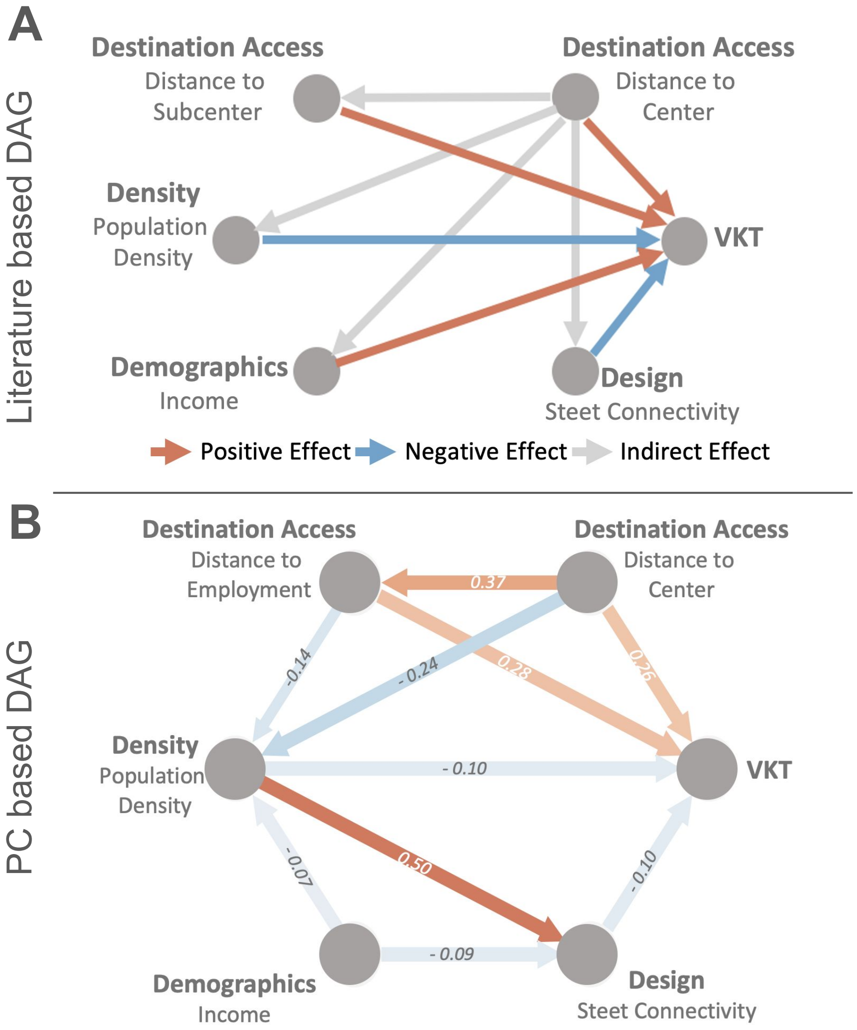
3.1 Causal urban form effects partially confirm previous assumptions

Similarities:

- Direct effects of density, design and distance to employment and center on VKT
- Indirect effect of distance to center on distance to employment and density

Differences:

- no significant effect of demographics on VKT
- indirect effects between demographics, density and design



2. Methods

2.1 Data Prep and Feature Engineering

Data:

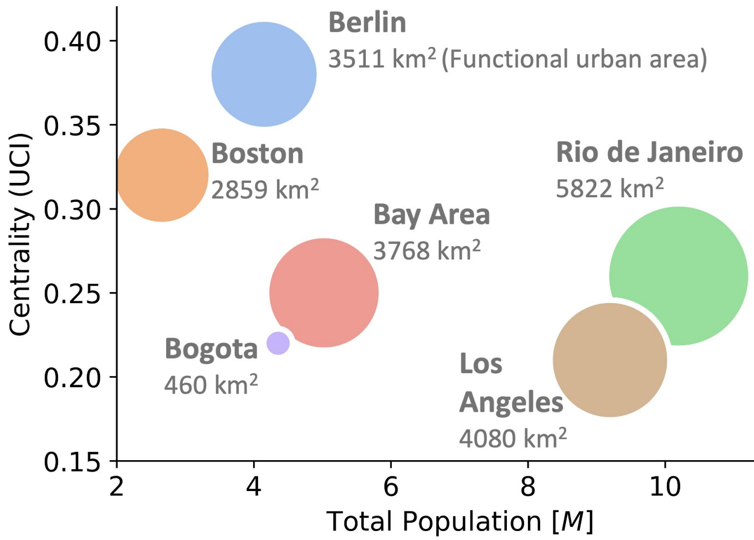
- **Travel distances:** Call Detail Records and GPS data
- **BE:** OpenStreetMap, Google Maps, local surveys, census

Cleaning:

- only commuting trips (6-10am)
- only Traffic Assignment Zones (TAZ) with > 10 trips
- only trips with origin and destination within Functional Urban Area

Target & Features:

- target: mean travel distance & mean emissions per TAZ
- features: Destination access, Density, Design, Demographics



D-Variable	Feature	Description
Destination Acessibility	Distance to city center	Distance from TAZ center to main city center.
Destination Acessibility	Distance to employment	Weighted average distance from TAZ center to 1% of all jobs.
Density	Population density	Number of inhabitants divided by area of TAZ.
Demographics	Income	Average household income per TAZ.
Design	Street connectivity	Number of intersections divided by area of TAZ.

Table 1. Predictive features of urban form based on the 6Ds of compact development

3. Results

3.2 Trends generalize across cities but differ in magnitude

Model results:

- varying generalization performance; tendency towards better predictions in more monocentric cities

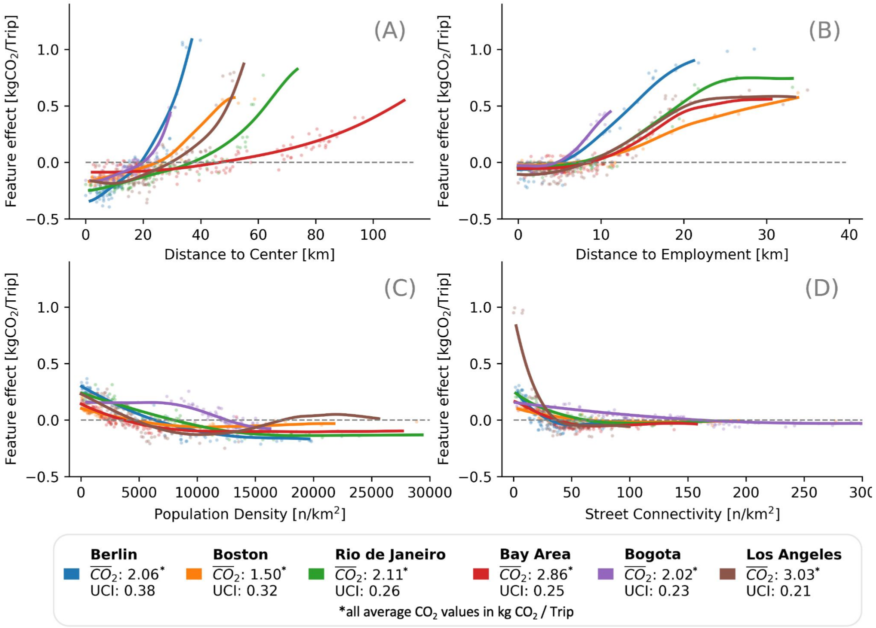
Feature effects:

- Distance to center and jobs has larger effects than density and street connectivity across cities
- very low densities and street connectivities should be avoided

Outliers:

- long tail for distance to center in sfo
- increasing effects for higher densities in Bogota and LA
- strong increasing effects for very low connectivities

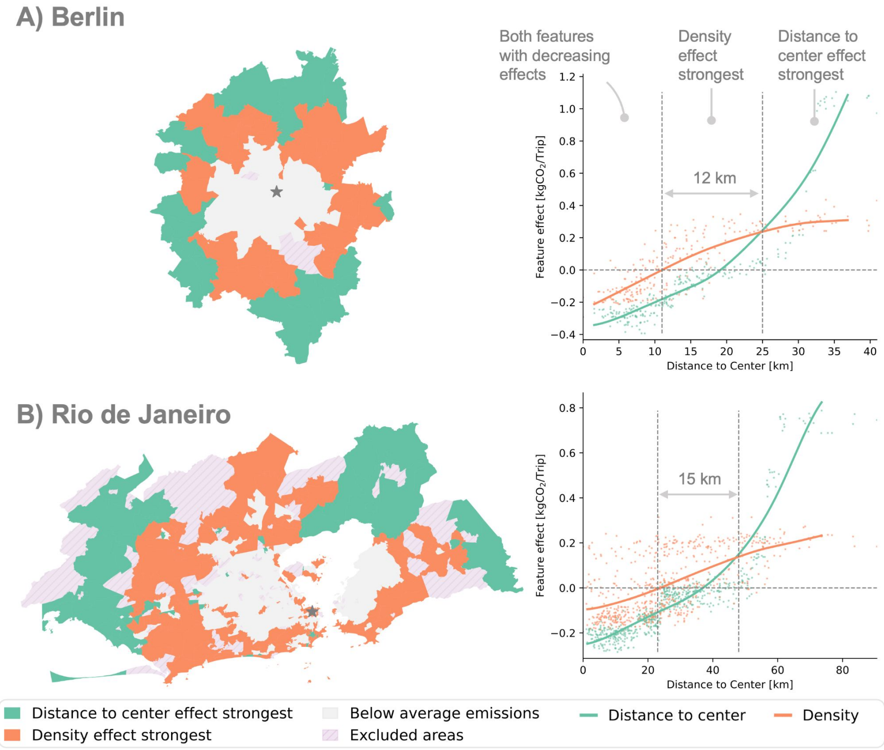
Metric	Berlin	Boston	Rio	Bay Area	Bogota	LA
R2	0.84	0.62	0.41	0.26	0.51	0.21



3. Results

3.3 Which urban form effect matters most depends on specific locations within cities

- comparison of density and distance to center for all locations with above city-mean emissions
- in contrast to previous work, we find in all cities specific locations, where density effect > distance to center effect
- f.e. in Berlin we find 12 km and in Rio de Janeiro a 15 km buffer zone



4. Discussion, Conclusion & Outlook

4.1 Takeaways

Accessibility to the main center is key (Fig 2)

- allocate new housing as close to center as possible
- avoid car trips at the very outskirts (mode shifts, occupancy, avoidance)

Improve access to jobs in peripheries (Fig 2)

- additional employment opportunities in outskirts can reduce VKT
- come at the risk of inducing new travel - additional measures needed

Prioritize density over accessibility at city-specific buffer zones (Fig 3)

- secondary urban centers have the potential to reduce trip distances

4. Discussion, Conclusion & Outlook

4.3 Conclusion next steps

Our results are a first step towards using big data & causal based approaches to help to translate national-scale scenarios for climate change mitigation from the IPCC to local-level recommendations. Yet, a lot of future work is required.

More spatially explicit analysis required

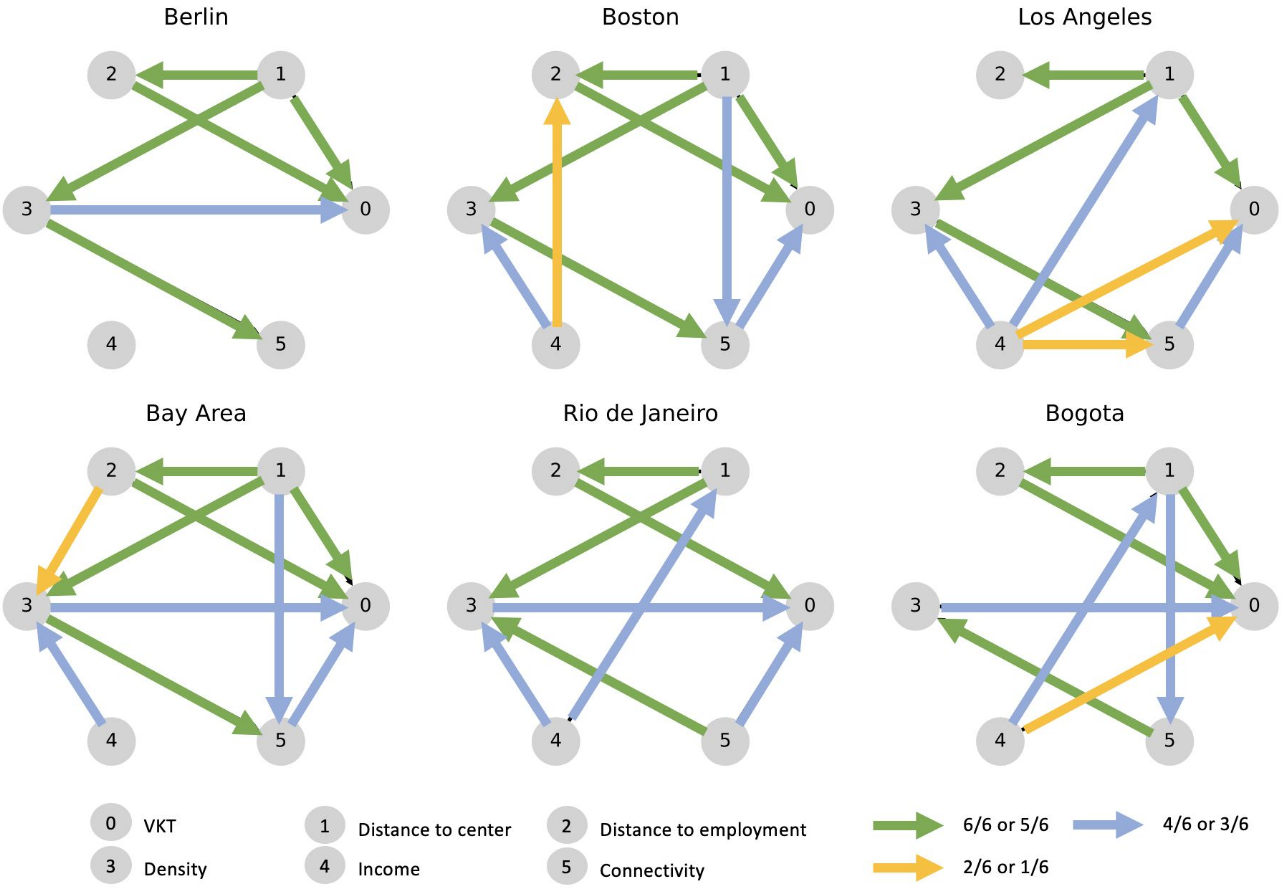
- differences in mono- vs. polycentric cities require more analysis - potentially using additional features
- differences in fast growing- vs. mature cities require more analysis

More causal approaches in context of urban science

- better representation of socio-demographics and attitude-based residential self-selection effects
- possibility to utilize DAG for causal inference approaches
- high potential to improve evidence-based policy-making

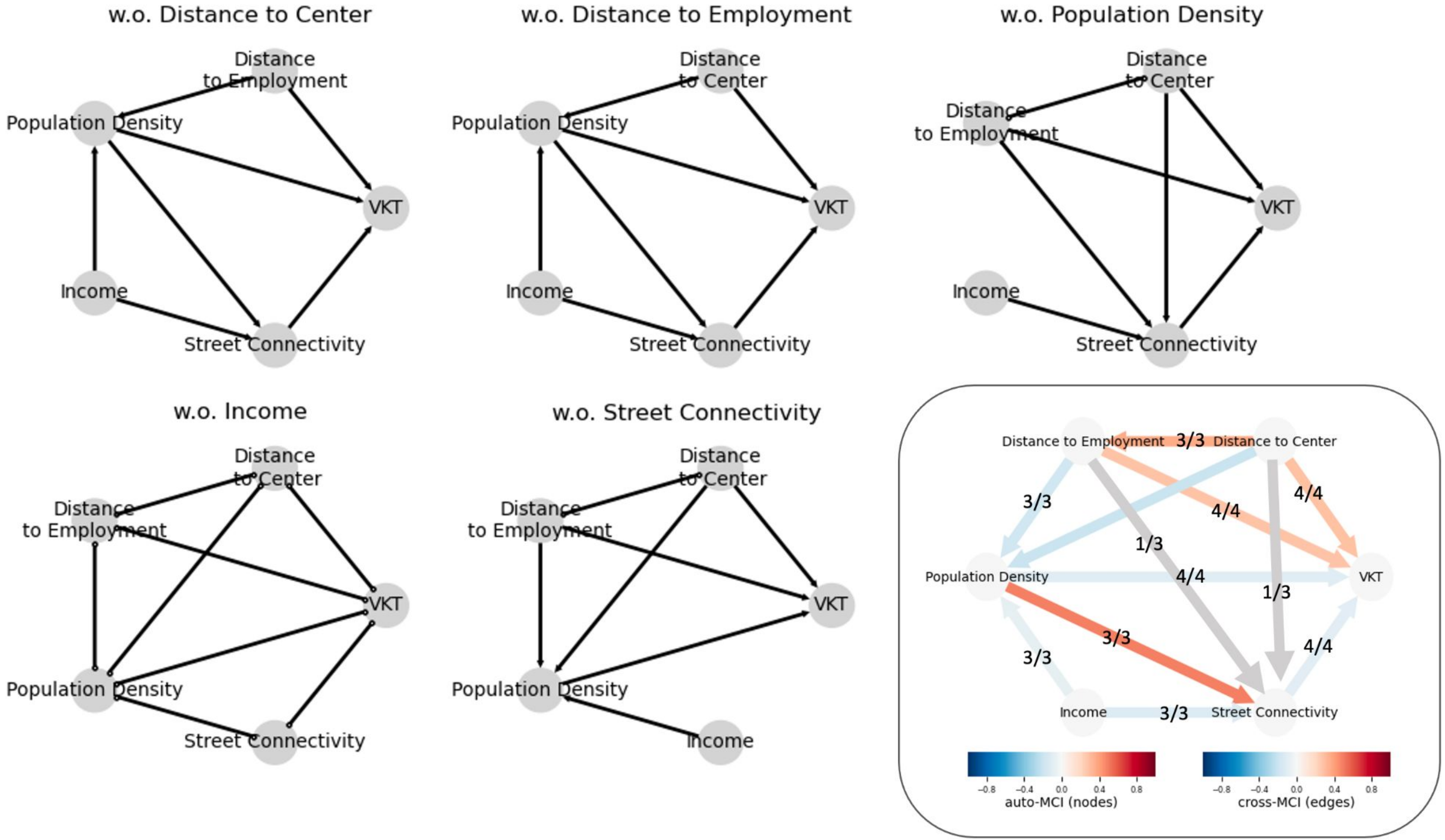
Appendix

Individual city analysis



Appendix

Leave one out analysis



Appendix

Results with CMIknn

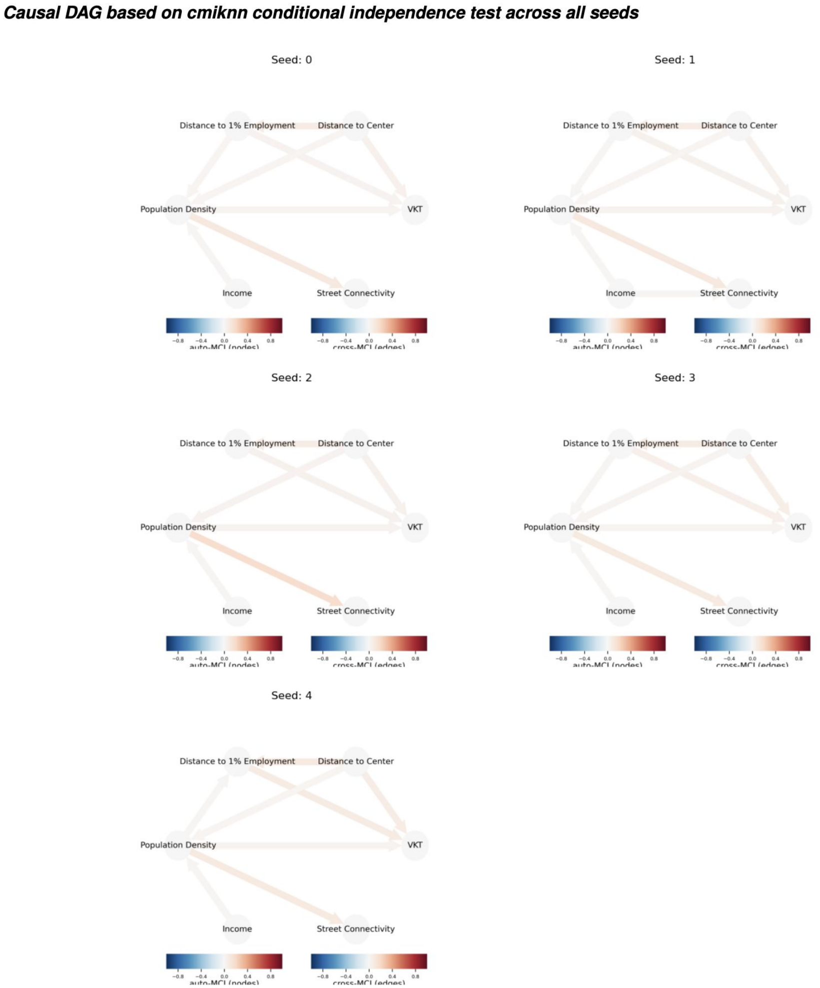


Figure 5. Causal DAG based on cmiknn conditional independence test across five seeds.