Master's project: A global typology for cross-city learning

Bastien Amez-Droz[1*]

[1]Human-Environment Relations in Urban Systems, École polytechnique fédérale de Lausanne, Station 2, Lausanne, 1015, Switzerland.

Corresponding author(s). E-mail(s): bastien.amez-droz@epfl.ch;

Abstract

The IPCC SR on Cities which will be compiled during the upcoming seventh assessment cycle will focus on climate solutions for different types of cities. However, a thorough and up-to-date underpinning of what different categories of cities may learn from each other is missing. Here, I bridge this research gap by developing a novel global typology of cities using Deep Embedded Clustering (DEC), grouping urban areas based on key socio-economic, infrastructural, and environmental attributes. A major contribution of this typology is the integration of high-resolution ODIAC CO_2 emissions data and OpenStreetMap (OSM) road network data, enabling a more detailed characterization of urban carbon footprints and infrastructure disparities. Then, by linking the city groups to research topics available from a recently developed database of the city-specific case study literature, I can paint a detailed picture of the cross-city learning potential. My results show that although 3,000 cities with fast population growth have not been covered in city-specific case studies, there are almost 1800 similar cities that have more than one study, particularly related to management practices and emissions, that these cities may learn from. By contrast, an albeit smaller group containing cities with fast-growing emissions are facing knowledge gaps, and consequently a low learning potential regarding research on emissions. More established cities with lower emission growth, where scientific knowledge about emission reductions is arguably less urgently needed, have large learning potentials.

Keywords: cities, typology, climate change, CO_2 emissions, cross-city learning, IPCC

EPFL Supervisors: prof Claudia Binder and Dr Simon Montfort
External Supervisor: Dr Tim Repke (Potsdam Institute for Climate Impact Research)

Main

Cities, where about 57% of the world's population live[1], play a central role in the global climate challenge, estimated to contribute by 70% of total emissions[2]. Despite their significance, research has disproportionately focused on large metropolitan areas, leaving smaller cities with under 250'000 inhabitants[1], home to 35% of the global population, understudied. This knowledge gap is particularly concerning for fast-growing urban areas, which are at a pivotal moment where present decisions will shape their future sustainability and avoid locking in carbon-intensive consumption models. As the primary places where people experience climate change firsthand and implement adaptation measures, cities serve as critical arenas for climate action. For instance, nature-based solutions, such as urban green spaces, wetland restoration, and sustainable drainage systems, are increasingly recognized and researched as effective strategies for enhancing climate resilience[3, 4] while providing co-benefits for biodiversity and human well-being. However, the ability of cities to learn from each other remains underexplored, limiting the transfer of effective adaptation and mitigation strategies[5]. Addressing this gap, this study seeks to answer: *What is the cross-city learning potential from climate change adaptation and mitigation evidence?*

Recognizing that the range of applicable climate solutions may vary depending on the local circumstances, the IPCC special report on cities[1] will focus on climate solutions adequate to urban types. In this context, possibilities for cities to learn from oneanother is of crucial importance to close evidence gaps. However, there is still little research that systematically quantifies how cities can learn from existing scientific evidence, especially in the case of small and rapidly growing cities in the Global South[5]. Recent research has shown that the evidence of case studies from these types of cities is scarce and also not well synthesized[6]. At the same time, the literature on urban climate solutions is expanding rapidly, making it increasingly difficult to keep track of emerging knowledge[7, 8]. Furthermore, this growing body of research is highly fragmented across disciplines, from urban planning and environmental science to public health and engineering, which complicates access and synthesis of relevant findings. Systematic evidence synthesis is foundational as it provides reliable and robust methods to aggregate and make accessible existing knowledge[9, 10]. Existing typologies help address data gaps, yet a comprehensive assessment of cross-city learning potential from documented climate solutions is still missing. Chapter 5 of the IPCC Special Report[1] highlights the urgent need for improved typologies to enhance knowledge transfer, further reinforcing the motivation for this study.

To bridge these research gaps, this study develops a global, data-driven typology covering 10,328 small and medium-sized cities with populations between 50,000 and 500,000 inhabitants[11] and links it to the available city-specific case study evidence. The global data-driven typology creates clusters of cities with similarities within the group and disparities across, allowing for structured comparisons and analysis. By integrating key city-level indicators, including population size, gross domestic product, growth rates, greenness, built-up area, heatwaves magnitude[12], road infrastructure[13], and emissions data[14], this typology captures critical factors shaping climate action. A major contribution lies in attributing gridded emissions and road network data to each individual city across the globe. CO_2 emissions data is essential for understanding the urban contribution to global climate change and assessing the effectiveness of mitigation strategies. Road network data is critical as it directly influences transportation patterns, accessibility, and infrastructure development, all of which significantly impact a city's carbon footprint and climate resilience. I link the available city case studies developed in ref [6], containing the results from a structural topic model (STM) and city-specific case studies attributed to individual cities. This allows me to show which topics are prevalent for which types of cities and what evidence cities may potentially learn from.

The methodological approach follows a structured three-step procedure. First, a typology of cities is developed using Deep Embedded Clustering (DEC), a state-of-the-art clustering method that outperforms traditional algorithm *k-means* with a large margin[15]. This classification identifies 5 city groups based on key variables, ensuring a robust representation of urban diversity. Second, scientific evidence on climate solutions is systematically linked to individual cities[6]. This step involves visualizing the geographic distribution of clusters while identifying the key characteristics that define each group. Research topics related to climate change adaptation and mitigation are selected, and their presence within each cluster is analyzed. The topic model is built using OpenAlex, the largest open-source scientific database[16]. It encompasses 93% of relevant IPCC references from AR6, ensuring comprehensive coverage of documented solutions. Third, the insights from both steps are synthesized to create the most extensive systematic map and database of climate solutions by city type, spanning over 20,000 case studies worldwide. This approach enables a more targeted cross-city learning process, facilitating the identification of relevant climate strategies for under-researched cities.

Results

Typology of small and medium-sized cities reveal distinct challenges across the world

I identify 5 distinct types of cities using key covariates like population, CO_2 emissions, GDP, road infrastructure, and the data-driven deep embedded clustering algorithm (DEC). The clustering results are not perfectly defined into clear and easily explainable categories due to the nature of unsupervised learning. Since the algorithm groups cities based on similarities in multiple dimensions without predefined labels, some cities may not fit neatly into a single category. Instead of rigid classifications, the focus is on identifying the dominant tendencies within each cluster.

The clustering results in Fig. 1a reveal a strong regional segregation, reflecting economic and infrastructural disparities between the Global North and the Global South. Clusters 1 and 2, primarily composed of cities in the Global South, include rapidly growing urban centres in Sub-Saharan Africa, South and Southeast Asia, and parts of Latin America. Cluster 1 (Fast Growing Cities) represents the bulk of these, with cities in China, India, and Africa experiencing rapid economic and demographic expansion but only moderate CO_2 emissions growth. Despite their rapid urbanization, emerging decarbonization efforts — particularly in the transport sector — may be mitigating emissions growth. However, these cities face major challenges related to underdeveloped infrastructure, including limited road networks and inadequate public services, which could hinder sustainable development. Managing urbanization sustainably while maintaining economic momentum remains their key challenge, with some cities following smart city models to improve resilience.

Cluster 2 (High Emission Growth Cities), found mainly in Southeast Asia, parts of the Middle East, and tropical regions, exhibits the highest CO_2 emissions growth, often linked to carbon-intensive industrialization and weak regulatory frameworks. These cities are typically smaller but experience disproportionate emissions increases without equivalent economic gains. Many are located in forested or coastal regions, making them greener than other clusters but also highly vulnerable to extreme heat events. Their reliance on fossil fuels and limited infrastructure—characterized by a lack of motorways and low built-up density—makes transitioning to a sustainable urban model particularly challenging. A major concern is their dependence on energy-intensive cooling solutions — exacerbating their emissions — and urban heat island effects.

In contrast, Clusters 3 and 5, predominantly in the Global North, encompass well-developed urban centers with more established infrastructure. Cluster 3 (Major Developed Cities) includes large cities in North America, Western Europe, Japan, and Australia, which are historically high emitters but now experiencing emissions stabilization or decline. These cities feature extensive motorway networks, the highest built-up density, and a large share of asphalted roads. However, they face challenges in transitioning toward sustainability due to entrenched infrastructure and spatial constraints that complicate the adoption of low-carbon transport and building solutions. With relatively low population and GDP growth, maintaining economic leadership while implementing ambitious climate policies is a primary concern.

Cluster 5 (Medium-Sized Cities with Intermediary Growth and Emissions) is more geographically diverse, including cities in developed regions such as Europe and North America but also wealthier urban centers in the Global South, such as those in Argentina, South Africa, and coastal China. These cities exhibit high CO_2 emissions and GDP per capita, with relatively well-developed infrastructure, including asphalted roads and some motorways. On the other hand, their growth rates are moderate meaning that they're are achieving a stability. The key challenge for these cities is balancing economic stability with environmental responsibility, particularly in emerging economies from this cluster where policy direction may fluctuate between economic expansion and sustainability commitments with emissions reduction.

Cluster 4 (Low CO_2 High GDP Growth Cities), though mostly found in poorer countries, also includes some cities in the UK, Spain, and Italy, highlighting a mix of urban contexts with relatively low emissions but strong economic growth. These cities share similarities with Fast Growing Cities but tend to be smaller and have lower GDP and emissions overall. However, their CO_2 emissions are increasing at a faster rate than those in the Fast Growing Cities cluster, suggesting they are at an

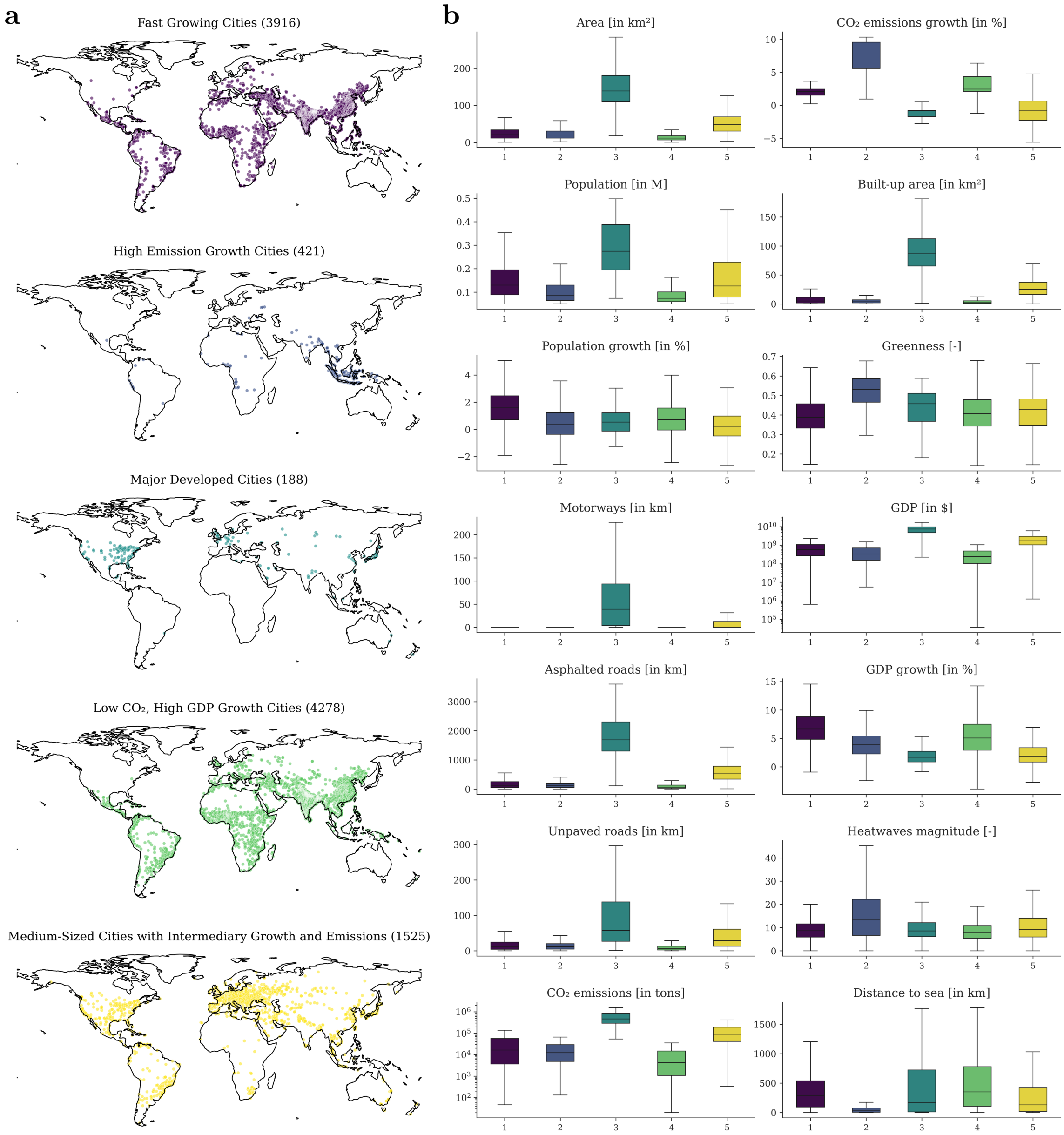


Fig. 1 – Overview of the typology. Panel a. shows the repartition of each city group across the globe. Each dot depicts one urban centre. Panel b displays the distribution of the values for each covariate across each city group. Absolute CO_2 emssions and GDP are showed in a log scale to imporve readability. Colors and order are consistent between panels a and b.

earlier stage of industrialization. With the lowest CO_2 emissions per capita and underdeveloped road networks, these cities must navigate the challenge of sustaining economic growth without adopting a high-carbon development trajectory.

This typology underscores the stark contrast between Global North cities, which are focusing on emissions reductions, and many Global South cities, which must balance rapid urbanization with growing environmental pressures. While developed cities grapple with decarbonization and infrastructure adaptation, rapidly growing cities in the Global South face pressing challenges related to uncontrolled expansion, infrastructure deficits, and climate vulnerability. The disparities in infrastructure, economic power, and policy priorities highlight the need for tailored sustainability strategies that address the unique challenges faced by each cluster.

Vast cross-city learning potential from scientific evidence

To investigate the learning potential of similar cities from the existing scientific evidence, I match case studies to individual cities in the typology, allowing me to aggregate how many research topics are included in a given group (Fig. 2). The research topics in the literature span across a wide range of topics (Fig. 3), covering the impacts of environmental changes, adaptation strategies, mitigation efforts, and cross-cutting governance and policy considerations. While impact studies do not directly mitigate climate change, they enhance understanding of problem exposure, which is essential for developing effective solutions.

The extent to which cities may learn from each other varies considerably. Cities with fast population growth (cluster 1), totaling around 5,000 cities across the globe (Fig. 2), may learn from nearly 1,800 studies, most of which cover climate change mitigation and CO_2 emissions and management-based adaptation. Management-based adaptation refers to implementing changes in practices or policies to address specific challenges[17, 18]. For example, a city facing increased flooding due to climate change might revise its zoning laws to prevent construction in flood-prone areas[19] and invest in better stormwater management systems, like during the reconstruction of New Orleans after hurricane Katrina[20].

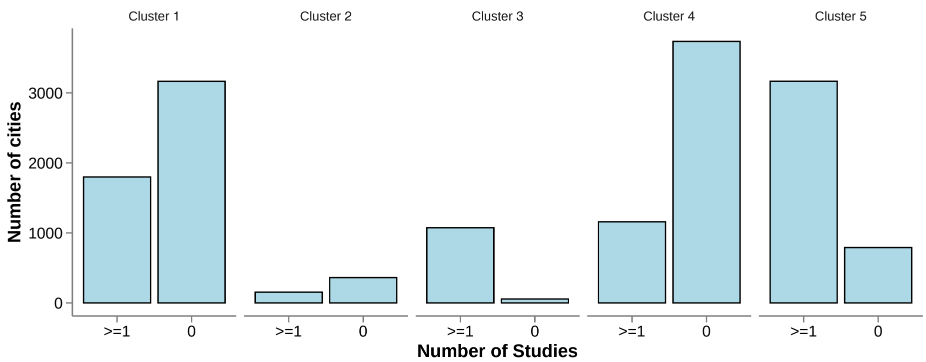


Fig. 2 – Learning potential per group The barplot shows, per cluster, how many cities are covered by one or more publications and how many aren't covered at all.

For the smallest group of cities, such as the low CO_2 - high GDP growth (cluster 4), the proportion of researched cities is only 33%. However, since this group consists of a great number of urban centers, the sheer quantity of cities sharing a similar type still provides a substantial amount of knowledge to be shared, with over 1,000 studied cities.

By contrast, my results show that medium-sized cities with intermediary populations (cluster 5) and emission growth, mostly situated in the developed world, face the best basis for evidence-based decision-making. Yet, these cities are at the same time best covered in the literature. Past research has shown that the larger and more developed a city is, the better it is covered in city-specific case studies.[5, 6, 21] Specifically, in this cluster, around 700 cities are not covered in any city-specific case study, but more than 3,000 cities are covered by more than one case study, thus exhibiting a high potential for under-researched cities to learn from. The evidence to take insight from is concentrated particularly around typical urban climate change mitigation topics for development, such as emissions, transport, buildings, and energy. Furthermore, governance and policy studies are needed for big cities in cluster 3, as their size, density, and economically well-established nature require systematic changes rather than new development paths to follow. Like those in cluster 5, these cities are also in the process of rethinking their urban transport systems.

Characterized by high-emission growth, cluster 2 presents a very low overall learning potential. First, there are little more than 100 studies available overall, covering only about 120 cities. Second, despite their key challenge being their fastest overall emission growth (see Fig. 1b), only four studies have been identified with this subject as their main topic. Hence, future research should investigate in more detail mitigation strategies for cities with fast emission growth, such as those in this group. Given the complexity of emissions problems, simulation-based research approaches are particularly needed. On the other hand, this cluster does have some learning potential regarding the impact of its

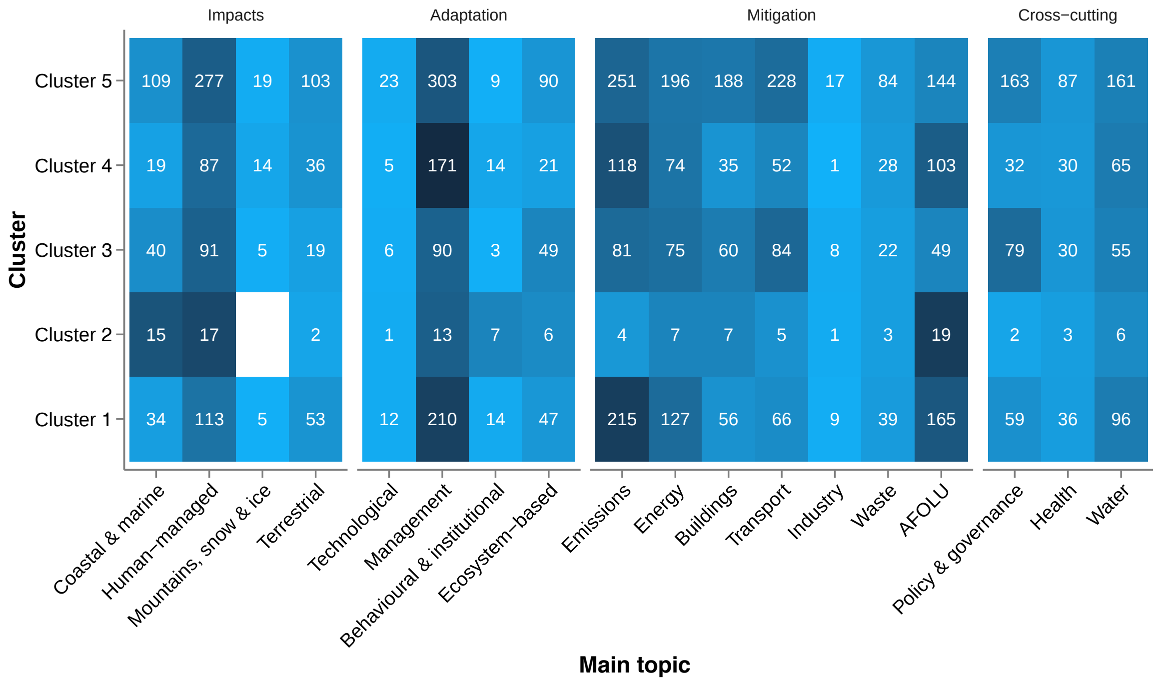


Fig. 3 – Size of research topics in the entire literature on cities by group. The grid gives the number of studies per city and topic. A darker shade of blue signifies a higher learning potential for other cities across the cluster. AFOLU stands for "Agriculture, Forestry and Other Land Use".

coastal and marine environment, with 17 studies covering topics like vulnerabilities of cities to sea-level rise, storm surges, and marine ecosystem shifts. However, nature-based solutions for urban heat adaptation seem to be lacking with only 6 ecosystem-based adaptation publications. While further research should prioritize these cases, cities in this group may already learn from neighboring urban areas in similar clusters, such as the low CO_2, high GDP growth cities, which benefit from 118 studies on the subject of emissions mitigaton.

Cluster 3, being mainly located in the OECD and consisting of major cities in terms of size and economic importance, presents a lower learning potential. These cities already benefit from extensive research, making them less dependent on external learning from similar urban areas. Their primary challenges are more related to implementing and scaling existing knowledge rather than generating new insights from comparable cases. Policy and governance studies remain essential in these cities, particularly for systematic changes in urban planning, infrastructure, and transportation systems.

Discussion: Bridging knowledge gaps for global urban climate action

This study highlights the vast learning potential across cities worldwide by identifying clusters of urban environments based on shared characteristics. The findings reveal both expected and unexpected links between cities across different countries, continents, and city sizes, demonstrating that urban challenges—and their solutions—often transcend geographical boundaries. For example, fast-growing cities in Africa and South Asia may find valuable insights in the experiences of rapidly urbanizing regions in Latin America, while mid-sized cities in Europe may have more in common with their North American counterparts than with their own capital cities. Recognizing these links is crucial for fostering cross-city learning and improving climate policy outcomes.

A key implication of this research is the identification of critical gaps in scientific evidence, particularly for cities experiencing the fastest emissions growth. Cities in cluster 4, characterized by high CO_2 emissions growth, have virtually no targeted research available in the literature. This is particularly concerning given that the IPCC has repeatedly emphasized emissions mitigation as a priority[1, 22, 23], yet the cities contributing most to future emissions remain severely underrepresented in scientific studies. Without sufficient evidence, policymakers in these cities may be forced to make decisions in the absence of empirical guidance[24], increasing the risk of ineffective or counterproductive policies. Future research must urgently address this gap by investigating mitigation strategies specifically tailored to high-emission growth urban environments.

More broadly, the lack of case studies for certain clusters underscores a fundamental challenge in climate policy: not all cities rely on high-level global reports like those from the IPCC to inform their decisions[2]. While global frameworks provide essential guidance, they often do not translate directly into local policy action, being confronted to institutional barriers and limited knowledge of the economic opportunities of low carbon development[25]. Instead, cities require accessible, context-specific research that is directly applicable to their urban realities[26]. The more localized and relevant the evidence, the more likely it is to be integrated into actual decision-making processes. Expanding city-specific case studies is therefore a critical step toward ensuring informed, science-based policymaking at the municipal level.

This study's typology — which integrates emission and road data from 254,362,704 1x1km grid cells and 36,845,972 line strings respectively, using GIS methods — provides a structured foundation for future cross-city learning efforts. By systematically categorizing cities based on environmental and socio-economic characteristics, this approach enables researchers and policymakers to identify relevant case studies more efficiently, reducing redundancy in research efforts. With nearly 6,500 out of 10,000 small and medium-sized cities already covered by one or more studies, there is significant potential for under-researched cities to draw lessons from those that have already been studied. This could help close the evidence gap while making climate research more efficient by reducing duplicative efforts[5, 6].

However, there are important caveats to consider. While the current typology provides valuable insights, future refinements could enhance its utility. Incorporating data on housing structures, building volumes[27, 28], waste management, and public transport systems could enable more precise comparisons between cities and help extract specific solutions rather than broad thematic trends. Additionally, as cities evolve over time, they shift between clusters, following different trajectories of urbanization, economic development, and emissions growth[29]. Understanding these transitions is key to predicting future urban challenges and designing adaptive policies that anticipate rather than react to change.

One promising avenue for further research is the identification of representative examples from each cluster. By selecting a set of well-documented case study cities that serve as benchmarks for their respective groups, researchers can strike a balance between scientific robustness and practical relevance for decision-makers.[6] These cities could act as hubs of knowledge, allowing other cities within their cluster to benefit from their experiences and avoid common pitfalls, like in the rapid urbanization of Lagos that faced expansion-related challenges[30] or during the reconstruction of Port-au-Prince after the 2010 earthquake, where European Urban Projects for a sustainable city clashed with the context of this urban centre[31].

To conclude, the findings highlight several priority areas for future research. While past studies have already pointed to the underrepresentation of cities in the Global South and fast-emitting urban centers[6], further expansion of the typology to incorporate more urban characteristics could refine the classification of cities into more specialized types. This study takes a significant step toward bridging this research gap by integrating global emissions data and road network information. This refinement allows for further enhancement of the effectiveness of cross-city learning, advancing the transferability of research findings and ensuring that scientific research is better aligned with the needs of urban decision-makers worldwide, allowing for scientists to be more directly implied in city-level decisions[32, 33].

References

[1] Programme, U. N. H. S. *World Cities Report 2024: Cities and Climate Action* (United Nations, 2024). URL https://www.un-ilibrary.org/content/books/9789211065602.

[2] Intergovernmental Panel On Climate Change (Ipcc). *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* 1 edn (Cambridge University Press, 2023). URL https://www.cambridge.org/core/product/identifier/9781009325844/type/book.

[3] Hobbie, S. E. & Grimm, N. B. Nature-based approaches to managing climate change impacts in cities. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20190124 (2020). URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0124. Publisher: Royal Society.

[4] Calliari, E. *et al.* Building climate resilience through nature-based solutions in Europe: A review of enabling knowledge, finance and governance frameworks. *Climate Risk Management* **37**, 100450 (2022). URL https://www.sciencedirect.com/science/article/pii/S2212096322000572.

[5] Lamb, W. F., Creutzig, F., Callaghan, M. W. & Minx, J. C. Learning about urban climate solutions from case studies. *Nature Climate Change* **9**, 279–287 (2019). URL https://www.nature.com/articles/s41558-019-0440-x. Publisher: Nature Publishing Group.

[6] Montfort, S. *et al.* A Global Systematic Map and Database of Climate Change Research on Cities – Insufficient Research Covers Fast-Growing Cities (2024). URL https://www.researchsquare.com/article/rs-5092456/v1. ISSN: 2693-5015.

[7] Minx, J. C., Lamb, W. F., Callaghan, M. W., Bornmann, L. & Fuss, S. Fast growing research on negative emissions. *Environmental Research Letters* **12**, 035007 (2017). URL https://iiasa.dev.local/. Number: 3 Publisher: IOP Publishing.

[8] Rosenzweig, C. *et al.* (eds) *Climate Change and Cities: Second Assessment Report of the Urban Climate Change Research Network* (Cambridge University Press, 2018).

[9] Hincks, S., Carter, J. & Connelly, A. A new typology of climate change risk for European cities and regions: Principles and applications. *Global Environmental Change* **83**, 102767 (2023). URL https://www.sciencedirect.com/science/article/pii/S0959378023001334.

[10] Creutzig, F., Baiocchi, G., Bierkandt, R., Pichler, P.-P. & Seto, K. C. Global typology of urban energy use and potentials for an urbanization mitigation wedge. *Proceedings of the National Academy of Sciences* **112**, 6283–6288 (2015). URL https://www.pnas.org/doi/abs/10.1073/pnas.1315545112. Publisher: Proceedings of the National Academy of Sciences.

[11] Urban population by city size. URL https://www.oecd.org/en/data/indicators/urban-population-by-city-size.html.

[12] Rivero, I. M. *et al.* GHS-UCDB R2024A - GHS Urban Centre Database 2025 (2024). URL http://data.europa.eu/89h/1a338be6-7eaf-480c-9664-3a8ade88cbcd. Publisher: European Commission, Joint Research Centre (JRC).

[13] OpenStreetMap contributors. Planet dump retrieved from https://overpass-turbo.eu/ . https://www.openstreetmap.org (2025).

[14] Oda, T., Maksyutov, S. & Andres, R. J. The Open-source Data Inventory for Anthropogenic CO_2, version 2023 (ODIAC2023): a global monthly fossil fuel CO_2 gridded emissions data product for tracer transport simulations and surface flux inversions. *Earth System Science Data* **10**, 87–107 (2018). URL https://essd.copernicus.org/articles/10/87/2018/?utm_source=chatgpt.com. Publisher: Copernicus GmbH.

[15] Guo, X., Gao, L., Liu, X. & Yin, J. Improved Deep Embedded Clustering with Local Structure Preservation 1753–1759 (2017). URL https://www.ijcai.org/proceedings/2017/243.

[16] Priem, J., Piwowar, H. & Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv* (2022). URL https://arxiv.org/abs/2205.01833.

[17] Broome, J. in *AR5 Synthesis Report: Climate Change 2014 — IPCC* (Intergovernmental Panel on Climate Change, 2015). URL https://www.ipcc.ch/report/ar5/syr/. Topic 4: Adaptation and Mitigation.

[18] Lawler, J. J. Climate Change Adaptation Strategies for Resource Management and Conservation Planning. *Annals of the New York Academy of Sciences* **1162**, 79–98 (2009). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2009.04147.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2009.04147.x.

[19] Stevens, M. R., Song, Y. & Berke, P. R. New Urbanist developments in flood-prone areas: safe development, or safe development paradox? *Natural Hazards* **53**, 605–629 (2010). URL https://doi.org/10.1007/s11069-009-9450-8.

[20] Kates, R. W., Colten, C. E., Laska, S. & Leatherman, S. P. Reconstruction of New Orleans after Hurricane Katrina: A research perspective. *Proceedings of the National Academy of Sciences* **103**, 14653–14660 (2006). URL https://www.pnas.org/doi/abs/10.1073/pnas.0605726103. Publisher: Proceedings of the National Academy of Sciences.

[21] Birkmann, J., Welle, T., Solecki, W., Lwasa, S. & Garschagen, M. Boost resilience of small and mid-sized cities. *Nature* **537**, 605–608 (2016). URL https://www.nature.com/articles/537605a. Publisher: Nature Publishing Group.

[22] Dhakal, S. *et al.* *Emissions Trends and Drivers* (Cambridge University Press, Cambridge, UK and New York, NY, USA, 2022).

[23] C40 Cities, Arup & University of Leeds. The future of urban consumption in a 1.5°c world (2019). URL https://www.c40.org/researches/the-future-of-urban-consumption-in-a-1-5c-world.

[24] Wiek, A. & Binder, C. Solution spaces for decision-making—a sustainability assessment tool for city-regions. *Environmental Impact Assessment Review* **25**, 589–608 (2005). URL https://www.sciencedirect.com/science/article/pii/S0195925504001234.

[25] Simarmata, H. A., Dimastanto, A., Santoso, S. I. & Kalsuma, D. Institutional Barriers of Low Carbon Development Planning in Indonesian Small Cities. *Low Carbon Economy* **05**, 105 (2014). URL http://www.scirp.org/journal/PaperInformation.aspx?PaperID=49581&#abstract. Number: 03 Publisher: Scientific Research Publishing.

[26] Howarth, C. & Painter, J. Exploring the science–policy interface on climate change: The role of the IPCC in informing local decision-making in the UK. *Palgrave Communications* **2**, 1–12 (2016). URL https://www.nature.com/articles/palcomms201658. Publisher: Palgrave.

[27] Biljecki, F. & Chow, Y. S. Global Building Morphology Indicators. *Computers, Environment and Urban Systems* **95**, 101809 (2022). URL https://www.sciencedirect.com/science/article/pii/S0198971522000539.

[28] Dong, B. *et al.* A Global Building Occupant Behavior Database. *Scientific Data* **9**, 369 (2022). URL https://www.nature.com/articles/s41597-022-01475-3. Publisher: Nature Publishing Group.

[29] Creutzig, F. *et al.* Scoping gaps in current assessments of cities and climate change (2024). URL https://irep.ntu.ac.uk/id/eprint/51809/.

[30] Adedire, F. M. Peri-urban Expansion in Ikorodu, Lagos: Extent, Causes, Effects, and Policy Response. *Urban Forum* **29**, 259–275 (2018). URL https://doi.org/10.1007/s12132-018-9336-5.

[31] Redon, M. The Model's Limitations. What 'Urban Sustainability' for Port-au-Prince? European Urban Projects Put to the Test by the Haitian City. *European Spatial Research and Policy* **20**, 41–56 (2013). URL https://www.czasopisma.uni.lodz.pl/esrap/article/view/7299. Number: 2.

[32] McPhearson, T. *et al.* Scientists must have a say in the future of cities. *Nature* **538**, 165–166 (2016). URL https://www.nature.com/articles/538165a. Publisher: Nature Publishing Group.

[33] Scholz, R. W. & Binder, C. R. Environmental literacy in science and society: from knowledge to decisions (2011). URL https://books.google.com/books?hl=en&lr=&id=ylPW3T06zZcC&oi=fnd&pg=PR12&dq=info:_UswdmzNeiEJ:scholar.google.com&ots=GwNzrP0YHu&sig=wCVgqJFx8Gdn1gFWjqXxR1a9_NU. Publisher: Cambridge University Press.

[34] Florczyk, A. *et al.* GHS-UCDB R2019A - GHS Urban Centre Database 2015, multitemporal and multidimensional attributes. Dataset (2019). URL http://data.europa.eu/89h/

53473144-b88c-44bc-b4a3-4583ed1f547e.

[35] Corbane, C. *et al.* The grey-green divide: multi-temporal analysis of greenness across 10,000 urban centres derived from the Global Human Settlement Layer (GHSL). *International Journal of Digital Earth* (2020). URL https://www.tandfonline.com/doi/abs/10.1080/17538947.2018.1530311. Publisher: Taylor & Francis.

[36] Kummu, M., Taka, M. & Guillaume, J. H. A. Gridded global datasets for Gross Domestic Product and Human Development Index over 1990–2015. *Scientific Data* **5**, 180004 (2018). URL https://www.nature.com/articles/sdata20184. Publisher: Nature Publishing Group.

[37] Russo, S., Sillmann, J. & Fischer, E. M. Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters* **10**, 124003 (2015). URL https://dx.doi.org/10.1088/1748-9326/10/12/124003. Publisher: IOP Publishing.

[38] European Commission. Joint Research Centre. *GHSL data package 2019: public release GHS P2019.* (Publications Office, LU, 2019). URL https://data.europa.eu/doi/10.2760/290498.

[39] Ehrlich, D., Freire, S., Melchiorri, M. & Kemper, T. Open and Consistent Geospatial Data on Population Density, Built-Up and Settlements to Analyse Human Presence, Societal Impact and Sustainability: A Review of GHSL Applications. *Sustainability* **13**, 7851 (2021). URL https://www.mdpi.com/2071-1050/13/14/7851. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

[40] Crippa, M. *et al.* EDGAR_emissions_urban_areas (2021). URL http://data.europa.eu/89h/c0c49cd7-4a80-4a94-8c34-375289c12b2d. Publisher: European Commission, Joint Research Centre (JRC).

[41] Chen, J., Zhao, F., Zeng, N. & Oda, T. Comparing a global high-resolution downscaled fossil fuel CO2 emission dataset to local inventory-based estimates over 14 global cities. *Carbon Balance and Management* **15**, 9 (2020). URL https://doi.org/10.1186/s13021-020-00146-3.

[42] An, N. *et al.* Monitoring of Atmospheric Carbon Dioxide over Pakistan Using Satellite Dataset. *Remote Sensing* **14**, 5882 (2022). URL https://www.mdpi.com/2072-4292/14/22/5882. Number: 22 Publisher: Multidisciplinary Digital Publishing Institute.

[43] Ahn, D. Y., Goldberg, D. L., Coombes, T., Kleiman, G. & Anenberg, S. C. CO2 emissions from C40 cities: citywide emission inventories and comparisons with global gridded emission datasets. *Environmental Research Letters* **18**, 034032 (2023). URL https://dx.doi.org/10.1088/1748-9326/acbb91. Publisher: IOP Publishing.

[44] Yang, E. G. *et al.* Using Space-Based Observations and Lagrangian Modeling to Evaluate Urban Carbon Dioxide Emissions in the Middle East. *Journal of Geophysical Research: Atmospheres* **125**, e2019JD031922 (2020). URL https://onlinelibrary.wiley.com/doi/abs/10.1029/2019JD031922. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019JD031922.

[45] M., H. P., Rehman, M. Z., Dar, A. A. & Wangmo A., T. Forecasting CO2 Emissions in India: A Time Series Analysis Using ARIMA. *Processes* **12**, 2699 (2024). URL https://www.mdpi.com/2227-9717/12/12/2699. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[46] Fan, C., Yang, Y. & Mostafavi, A. Neural embeddings of urban big data reveal spatial structures in cities. *Humanities and Social Sciences Communications* **11**, 1–15 (2024). URL https://www.nature.com/articles/s41599-024-02917-6. Publisher: Palgrave.

[47] Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N. & Fei, S. Automated content analysis: addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution* **7**, 1262–1272 (2016). URL https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12602.

[48] Halterman, A. Mordecai: Full Text Geoparsing and Event Geocoding. *Journal of Open Source Software* **2**, 91 (2017). URL https://joss.theoj.org/papers/10.21105/joss.00091.

[49] Park, H., Kim, D.-H. & Chang, S. Research Trend Analysis on Smart City based on Structural Topic Modeling(STM). *Journal of Digital Contents Society* **20**, 1839–1846 (2019). URL http://www.dbpia.co.kr/Journal/ArticleDetail/NODE09215450.

Methods & Data

I used data analysis, GIS, and machine learning methods to gather data from various sources, combine them and create a basis for the typology. For the integrity of said typology, some urban centres were dropped and not considered in my analysis because NaN values and false zeros threatens the integrity of the applied machine learnign methods. It should also be noted that the global scale of the used data required multiple memory optimizations, API timeout workarounds and GPU exploitation for neural networks training.

Table 1 below provides with a full overview of the variables gathered for every city, while the acquisition and computation processes of each variable will be detailed in this section of the report.

Table 1 – Overview of variables, their units, timescales, data sources, and references.

Variable name	Unit	Data timescale	Data source	Reference
Area	[km^2]	2019	GHSL	Rivero et al., 2024 [12]
Population	[M inhabitants]	2015	GHSL	Florczyk et al., 2019 [34]
Population growth	[%]	2000-2015	GHSL (derived)	-
Motorways	[km]	2024	OSM (derived)	OSM contributors, 2024 [13]
Asphalted roads	[km]	2024	OSM (derived)	OSM contributors, 2024 [13]
Unpaved roads	[km]	2024	OSM (derived)	OSM contributors, 2024 [13]
CO_2 emissions	[tons]	2022	ODIAC (derived)	Oda et al., 2023 [14]
CO_2 emissions growth	[%]	2000-2022	ODIAC (derived)	-
Built-up area	[km^2]	2015	GHSL	Florczyk et al., 2019 [34]
Greenness	[-]	2014	GHSL	Corbane et al., 2018 [35]
GDP	[$]	2015	GHSL	Kummu et al., 2018 [36]
GDP growth	[%]	2000-2015	GHSL (derived)	-
Heatwaves magnitude	[-]	1980-2010	GHSL	Russo et al., 2014 [37]
Distance to sea	[km]	2024	GHSL (derived)	-

GHSL: Global Human Settlement Layer; OSM: OpenStreetMap; ODIAC: Open-source Data Inventory for Anthropogenic CO_2. The greenness and heatwave variables don't have a unit because they are an index. The data timescale indicates the year the value corresponds to or the time extent considered for the value computation. "Derived" means that the variable was acquired through some further processing steps.

Typology development

Data acquisition and processing.

Urban centres

The typology is built upon the Global Human Settlement Layer Urban Centre Database (GHS-UCDB) [38], a comprehensive open-source dataset that provides globally consistent information on urban centers. The latest version of the GHS-UCDB includes data on population, gross domestic product (GDP), green areas, built-up surfaces, and heat extremes, offering a solid foundation for city classification. One of its key strengths is its coverage of urban centers in the Global South, ensuring a more inclusive and representative analysis. Several key variables are directly extracted from the GHS-UCDB, including population size, GDP, green areas, built-up areas, and extreme heat exposure, which have been used in urban climate studies[39]. Additionally, some variables are derived from the dataset, such as population Compound Annual Growth Rate (CAGR) (1)), GDP CAGR, urban center area, and distance to the nearest coastline, calculated as the shortest distance between the centroid of each urban center and global seas.

$$\text{CAGR} = \left(\frac{V_f}{V_i}\right)^{\frac{1}{t}} - 1 \tag{1}$$

CO2 emissions

CO_2 emissions data is sourced from the Open-source Data Inventory for Anthropogenic CO_2 (ODIAC), a high-resolution dataset that is evidence driven with the use of night light data from the GOSAT mission, and offers finer spatial resolution compared to the Emissions Database for Global Atmospheric Research (EDGAR)[40]. ODIAC has been demonstrated to perform well at the city scale[41, 42], and its validation against EDGAR aligns with findings in existing literature[43, 44], with a global R^2 of 0.675 for the year 2015 between my extracted ODIAC values and EDGAR values from the GHSL-UC. Correlations by world region vary and can be observed in the figure A4. This dataset is particularly advantageous for assessing emissions in smaller urban settlements due to its granularity and was therefore chosen over EDGAR for this reason and its connection to evidence with satellite data.

To estimate emissions for each urban center, monthly ODIAC GeoTIFFs from 2000 to 2022 were processed using rasterio and rasterstats in Python, extracting emissions by overlaying raster data with urban center polygons. Those python packages allow large geographic data processing with local reading and raster pixel row by row handling. This gives our first emssions variable, the absolute total city CO2 production for the most recent year: 2022. The resulting time series data was used to compute emission trends via Autoregressive Integrated Moving Average (ARIMA) models, selected based on the Akaike Information Criterion (AIC). To assess short-term trends, I computed the average annual growth rate for the 5 years forecast (mention extended figures). The ARIMA models capture city-specific autoregressive dynamics, following state-of-the-art methodologies in emissions forecasting [45]. An overview of the whole extraction process can be found in the extended data figures A3. Overall, 921,604 1x1km gridded cells are extracted for each month, meaning 254,362,704 km^2 overall.

A challenge in the dataset is the presence of zero emissions values for certain urban centers in specific years. This stems from ODIAC's reliance on night time light data for nonpoint emission disaggregation, which may introduce biases, particularly in developing and least-developed countries [14]. While these values likely indicate very low rather than truly zero emissions, they are treated as zeros in the dataset. A five-year consecutive emissions record prior to 2022 (2018–2022) was required for trend estimation using ARIMA models, ensuring robust trend assessments despite potential data gaps. Urban centres with insufficient data were dropped (add number).

Roads

Road infrastructure data was extracted from OpenStreetMap (OSM)[13], an open-source dataset providing globally consistent and categorized road networks. OSM is particularly relevant for this study as it offers a single-source, structured categorization of road types. However, due to its participatory nature, coverage and accuracy may vary, particularly in less-developed regions. Some countries have highly detailed road networks, while others remain sparsely mapped, leading to potential disparities in infrastructure data quality. Examples in extended data figures show that (ref).

To extract road network data, we retrieved all roads within each urban center using the Overpass API, ensuring the use of the most recent dataset. Given the large data volume, preprocessing methods were implemented to efficiently handle and store road data. Road lengths were computed in meters, carefully adjusting for variations in latitude to maintain accuracy in distance calculations. The road infrastructure was categorized into three primary types:

- Motorways: High-capacity roads facilitating regional and international connectivity.
- Asphalted roads: Major and minor roads that form the backbone of intra-urban mobility.
- Tracks and unpaved roads: Low-infrastructure roads, typically indicative of informal or rural-urban transition zones.

This process can be viewed as a flowchart in the extended data figure A2 and lead to the extraction of 13.8 billion meters of relevant road segments.

Clustering methods, machine learning

For the typology construction, only small and medium-sized cities were considered, reducing the dataset from 13,115 to 10,328 urban centers within the 50,000 to 500,000 population range. This threshold is based on OECD classifications [cite] and focuses on cities with high learning potential while helping statistical robustness by avoiding large and megacities outliers.

The clustering was performed using Deep Embedded Clustering (DEC), which differs from traditional clustering methods by jointly learning feature representations and cluster assignments in a low-dimensional latent space. Unlike K-means, which operates directly in the original feature space, DEC trains an encoder-decoder neural network to first compress data into a lower-dimensional space and then iteratively refine cluster assignments. This approach has been effectively used for urban typology classification.[46] (citation not perfect, other embedding technique but same idea). [add sentence about dropping NaN and False zeros]

Before clustering, data was normalized using StandardScaler from sklearn, ensuring that all variables contribute equally to distance calculations in the latent space. The encoder-decoder structure follows a progressive densification approach with layers of $128 \rightarrow 96 \rightarrow 64$ neurons, ultimately reducing to a 5-dimensional latent space. The architecture was kept relatively shallow to allow the data to express itself more effectively. The model was trained for 200 epochs using Stochastic Gradient Descent (SGD) with a learning rate of 0.01 and momentum of 0.9, with a batch size of 32.

After encoding the data, K-means clustering was applied to the learned latent representations, with the optimal number of clusters determined using silhouette scores. The silhouette coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. A detailed silhouette score plot is provided in the extended data figure A6, demonstrating the elbow point used to identify the best number of clusters where the rate of decrease shifts. Based on this analysis, five clusters were selected as the optimal solution, ensuring a balance between compactness and separation of groups and vizualization capabilities.

Linking the data on scientific evidence to city groups

To support this analysis, I leverage an existing topic model[6] developed in previous research to systematically categorize climate change studies related to cities that extracts insights from the big litterature[47]. This model was constructed using a comprehensive literature dataset retrieved from OpenAlex, one of the largest open-source bibliographic databases. The dataset, queried in February 2023, includes a wide range of studies on climate change mitigation, adaptation, and related urban impacts.

The selection process of the literature began with a broad query in OpenAlex using climate change-related keywords. To refine the dataset and ensure relevance, a supervised machine learning classifier (XGBoost) was employed. This classifier was trained using a manually annotated sample of 1,020 research article abstracts, applying text preprocessing techniques such as tokenization, stopword removal, and frequency-based filtering. To mitigate class imbalance, upsampling was applied to the minority class. Model performance was optimized through hyperparameter tuning using a nested cross-validation framework, achieving an out-of-sample F1 score of 0.952.[6]

To extract city-level insights, the dataset was further processed using a combination of geoparsing techniques:

- Named-entity recognition (NER): A pre-trained geoparser, Mordecai[48], was used to identify and geolocate city names mentioned in unstructured text. This process involved transformer-based encoding to recognize place names and resolve toponym ambiguities.
- Keyword matching: An explicit search for city names from the GHSL-UCDB database[38] was performed, incorporating rules to handle non-unique city names through contextual keyword matching and probabilistic assignment.

Validation of these approaches involved manual annotation of 600 articles to evaluate prediction accuracy, yielding an optimal method combining both approaches (F1 score = 0.79). However, accuracy varied across regions, with lower performance in South America and Africa due to complex city naming conventions[6].

To classify studies into meaningful thematic categories, structural topic modeling (STM)[49], an unsupervised machine learning algorithm, was employed to extract latent topics from unstructured text. Multiple models were tested with topic numbers ranging from 60 to 240, with the final selection of 220 topics providing an optimal balance between granularity and interpretability[6]. The topics were then manually labeled and grouped into four overarching categories:

- Impacts (Coastal & marine, Human-managed, Terrestrial and Mountains, Snow & ice)
- Adaptation (Technological, Management, Behavioural & institutional and Ecosystem-based)
- Mitigation (Emissions, Energy, Buildings, Transport, Industry, Waste and AFOLU)
- Cross-cutting themes (Policy & governance, Health, Water)

Topics related to general research methods, geographical references, or non-substantive terms were excluded from the final analysis.

By utilizing this structured methodology, the topic model provides a robust framework for quantifying the distribution of studies across cities and topics, forming the basis of the analysis on cross-city learning and climate change research[6].

Declarations

Acknowledgements

I would like to thank prof Claudia Binder the opportunity of this master's project at HERUS.
I'm deeply indebted to Simon for his encouragement, his positivity and all the time he devoted to me that cannot be overestimated. Without his unconditional support throughout this project, I don't know how I would have done it.
I wish to thank Tim for his helpful advice regarding unsupervised models training and for accepting to be the external supervisor to this project. Many thanks to my fellow student Mateo for the discussions about our respective projects.
I'd also like to extend my thanks to my family and friends for their understanding when I had to pull long hours and for their moral support through small gestures. I'm especially extremely grateful to my parents for supporting me throughout my studies at EPFL.
Finally, I must thank my dog, Luna, for forcing me to take some needed breaks to go outside and enjoy a view of the forest and the lake, rather than my computer screen.

Conflict of interest/Competing interests

The author declares that no conflicts of interest exist.

Ethics approval

Not applicable.

Availability of data and materials/Code availability

All replication materials and data can be found on https://github.com/Krarouge/global_city_typology.

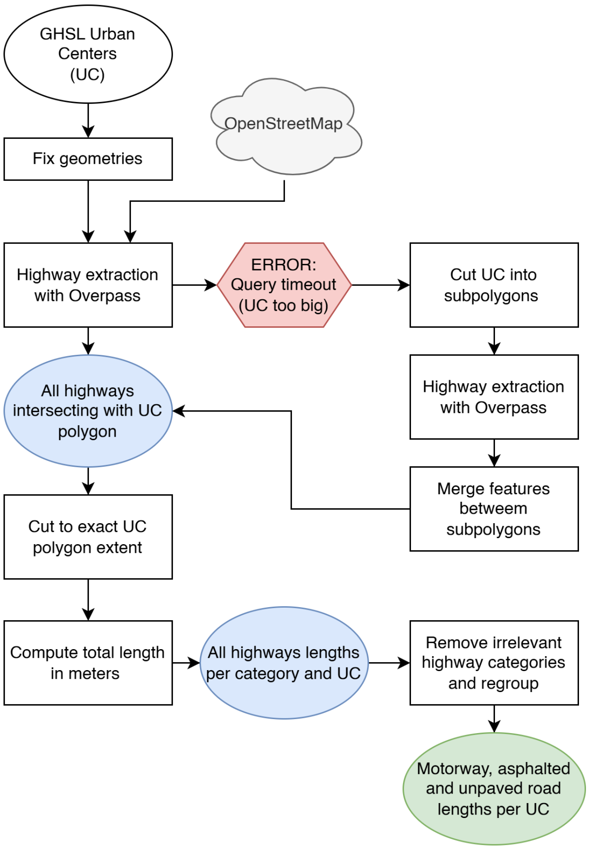
Appendix A Extended Data

a

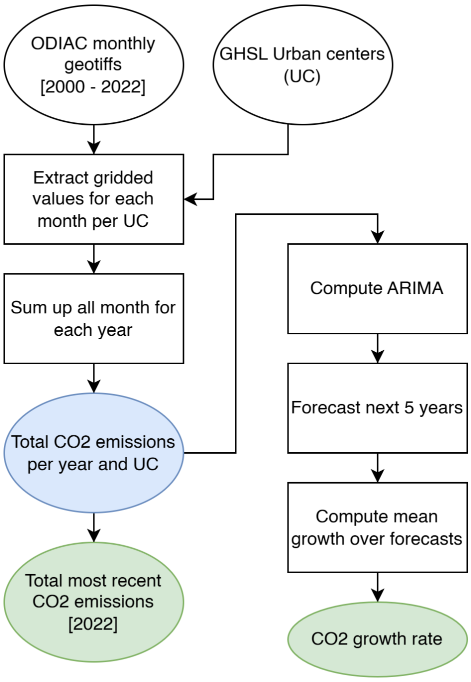
b

c

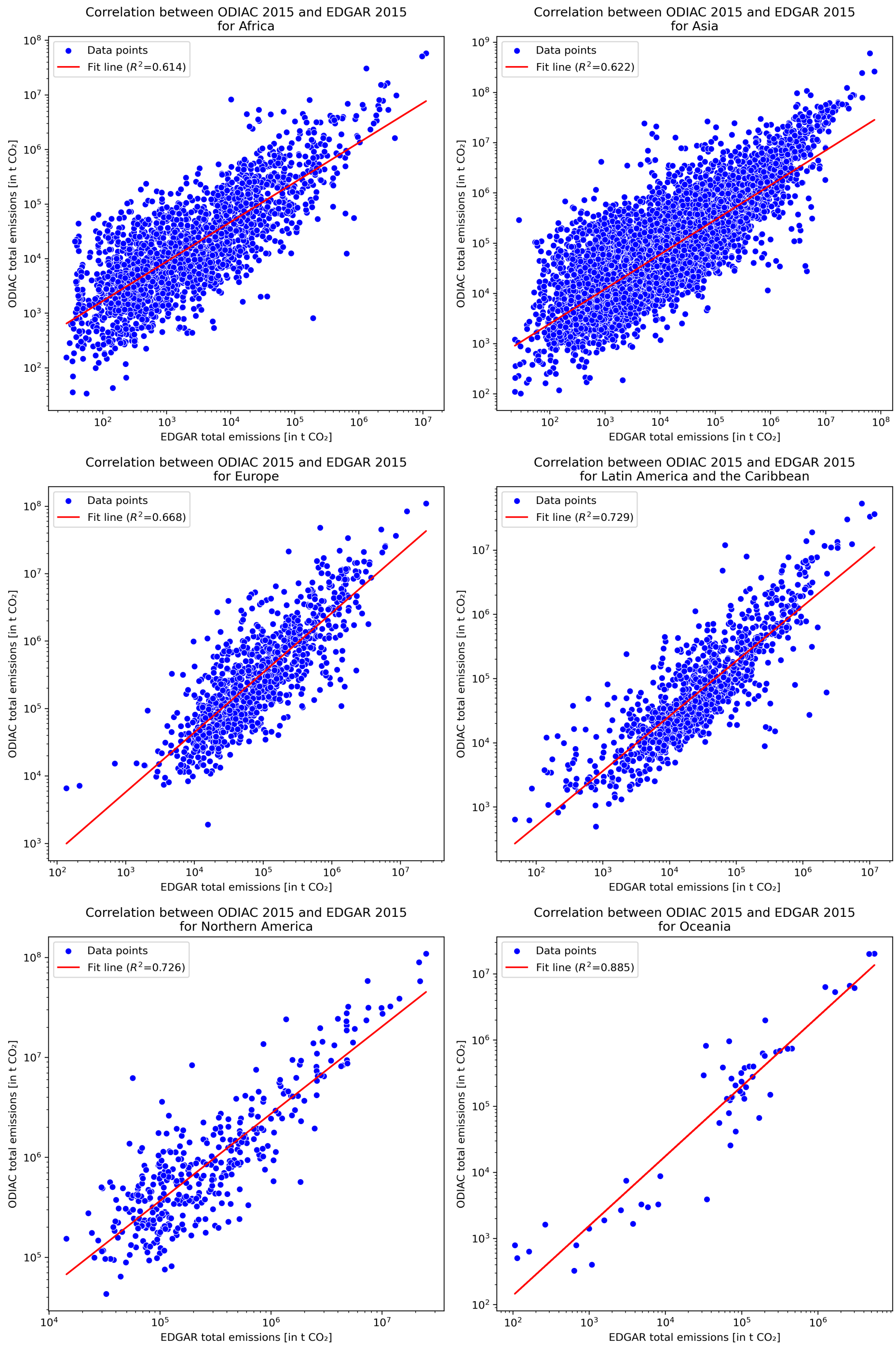
Extended Data Fig. A1 – OSM disparities across cities in the Global South. Being of participative nature, OSM data varies in quality across urban centres, particularly in less developped area such as Africa and Asia. The city of Kaolack in Senegal on picture a is a good example of an extensively covered road network, even for irregular tracks. Sumule in South Sudan on picture b is of similar size but has far less tracks covered by OSM. But paths on the image probably aren't meant for cars. Finally, image c is of Gobindpur in India, showing that even some track that are clearly used for motorised transportation aren't included in the OSM survey. Image data by Google ©2024.



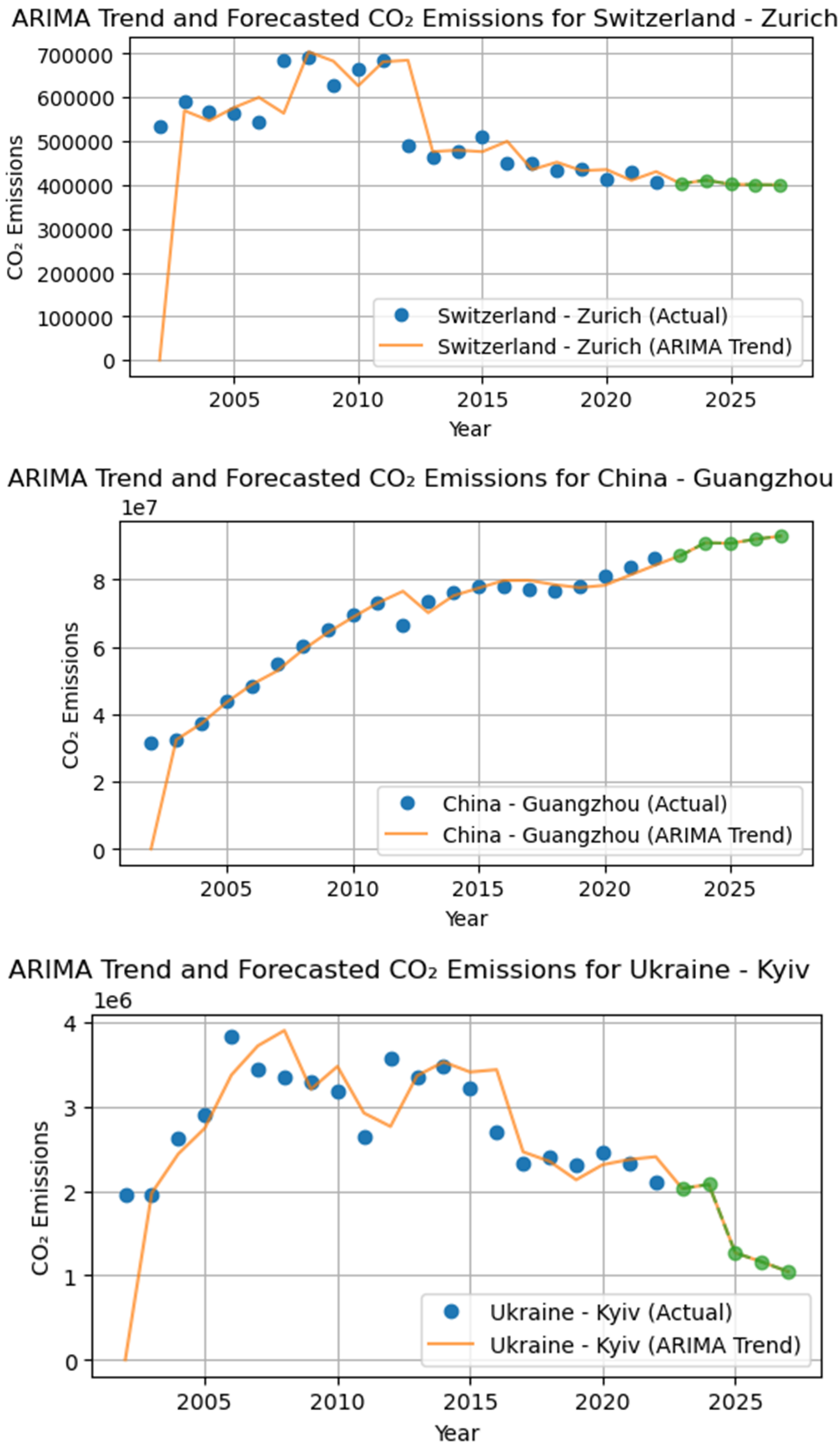
Extended Data Fig. A2 – OSM roads process. Intermediary products in blue and final product in green.



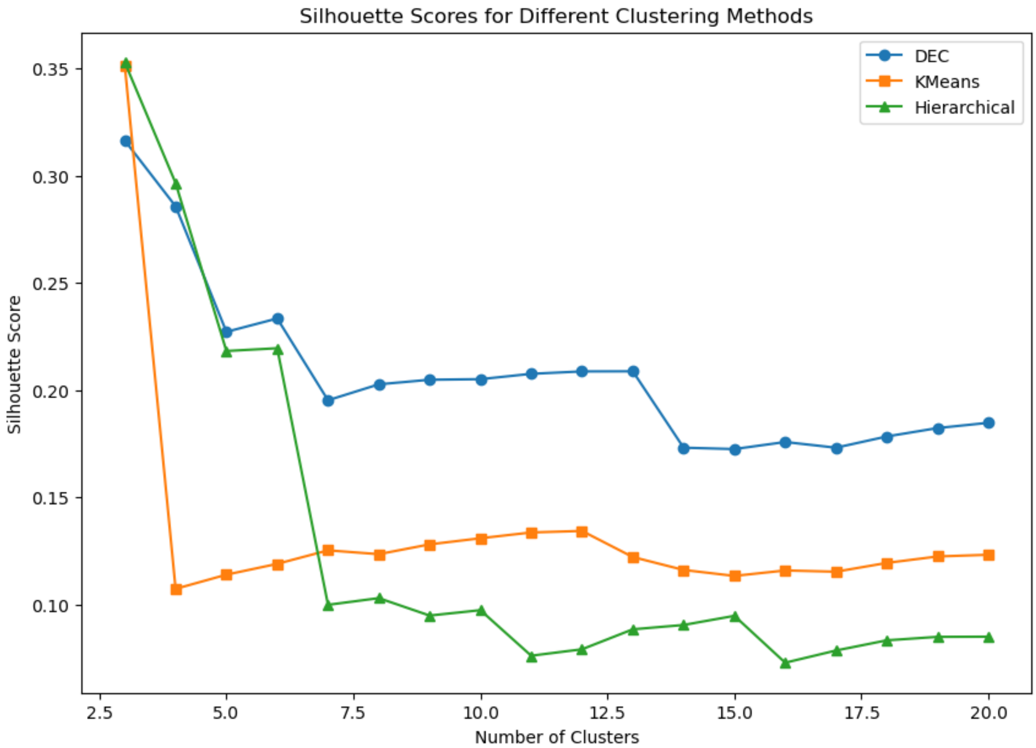
Extended Data Fig. A3 – CO2 process. Intermediary products in blue and final products in green.



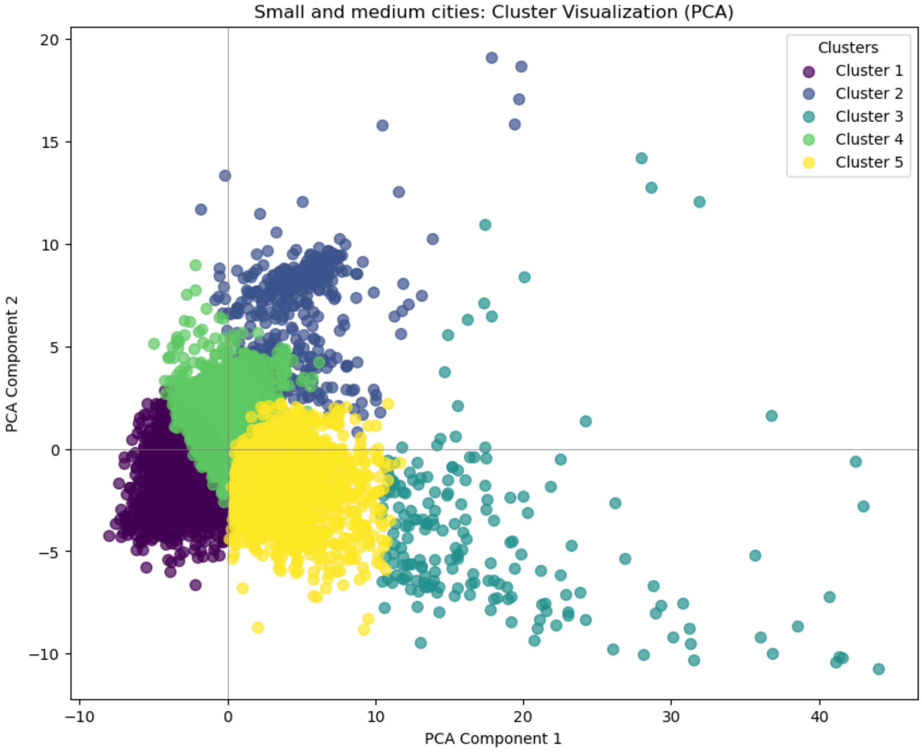
Extended Data Fig. A4 – Correlation between ODIAC and EDGAR datasets. Linear regression on EDGAR emissions data provided inside the GHSL urban centres database against ODIAC emissions computed over the urban centres polygons. Displayed on logarithmic scale and divided between world regions. The general correlation across all regions has an R^2 of 0.675.



Extended Data Fig. A5 – Computed ARIMA trends for sample cities. This figure provides with examples of the computed ARIMA trends using AutoARIMA with the Akaike information criterion. Here, Zürich is an example of slowly decreasing emissions, Guangzhou of increasing emissions and Kyiv of highly varying emissions, but decreasing. In green, the dots represent forecasted CO_2 emissions for the years 2023 to 2027. Emissions are in tons of CO_2.



Extended Data Fig. A6 – Silhouette scores of various clustering methods. The figure helps analyzing the performance of the different clustering methods with various number of clusters. In this case, 5 clusters is chosen because it represents an elbow before a stable plateau of silhouette score. 7 would also be a valid choice.



Extended Data Fig. A7 – Cluster vizualization through principal component analysis (PCA). The figure shows the dispersion of the cities variables across a PCA 2D space computed with the encoded variables after deep embedded clustering.