

DATA QUALITY ASSURANCE AND QUALITY CONTROL (QA/QC)

It is important to establish robust quality assurance and quality control procedures for cleaning your data. Make sure you have a document that clearly charts your QA/QC procedures. This means interrogating each step in the data collection and processing to identify and correct possible errors. This is important for ensuring your data is properly quality controlled. It also ensures transparency, by allowing others to trace the path of your data from raw to final which is important in making the data useful for policymaking and inspiring trust from policy-makers, researchers, and the public. Work with air quality experts and data scientists to ensure your data processing plan will generate the quality level required for your data objectives.

QA/QC procedures will vary for different instruments, pollutant species, and operating condition requirements. Because no two monitoring systems are alike (different instruments, monitoring methods, and different monitoring goals), data QA/QC will differ from project to project. However, key procedures in data QA/QC generally include:

- Data acquisition.
- Quality/parameter checks (automated and manual).
- System flags (common flags include invalid measurements, missing data, measurements below/above detection limit, zero/span checks, instrument repair period).
- Date and time stamp.
- Spatial coordinate stamp, if mobile (check for invalid or missing GPS stamp). As mentioned earlier in Section B, ensure that every record uses a common

time stamp that's based on a GPS satellite grade clock. You may need to apply time adjustments to the data, based on time of flight (when monitoring car takes off and begins data collection) and response time of sensor system.

- Other special flags, such as weather conditions that may affect measurements such as fog or extreme temperatures, and other anomalies and outliers.

Data QA/QC procedures should clearly articulate each step where data is processed (i.e. when data is redacted, adjusted, or transformed). For an example, see the summary of Breathe London project's data verification and quality assurance process on the website's Methodology page.

EDF has also developed a code repository for efficient data processing. Examples of open-source scripts we currently use can be accessed [here](#).

Potential pitfalls and other important factors:

- It is important to have clearly articulated data QA/QC procedures at the outset of the project *and* to identify the responsible party for the various steps. Without an effective QA/QC plan, you may end up with erroneous data that have to be discarded and can't be retrieved.
- The development of a data QA/QC plan can often be iterative. As data is collected and reviewed, you may find errors not previously captured and that additional procedures and flags are required.
- Make sure to review the data sufficiently often. Data QA/QC systems usually have automated components, however, manual checks are also critical. Make sure these are carried out regularly.

Implementation readiness checklist

- Project team assembled; external partners have signed MOU(s)
- Partners agree on roles, responsibilities, and processes for making decisions and identifying/resolving problems; project goals, data objectives, and appropriate data deliverables established
- Funding secured
- RFP(s) for monitoring system design, testing, installation, and management issued
- Instrument provider and/or specialist service provider retained
- Instrument systems tested and verified in the field at an acceptable level
- New monitors co-located with reference monitors; results show agreement
- SOPs for instrument operations (including a troubleshooting resource for drivers) developed and easily accessible for all relevant groups
- Data management system in place
- Data QA/QC protocol in place

Making sense of the data

In pursuing hyperlocal air pollution mapping — whether you deploy a large, dense, stationary monitoring network, or measure pollution with mobile platforms — you'll be managing a large amount of data. The 100x100 black carbon sensor network deployed in West Oakland collected more than 20 million lines of 1-minute data over the course of 100 days. The Houston mobile monitoring project generated almost 33 million valid pollutant data points over the course of nine months. Data processing of these large datasets involves, primarily, data aggregation and data reduction using specialized algorithms. Data analysis is then carried out to identify detailed spatiotemporal trends, or how air pollution levels vary over space and time.

The expertise and roles that are essential to analysing hyperlocal air pollution data include:



Data scientists, analysts and/or modelers

who ideally specialize in environmental data. These are professionals with the capabilities to analyze and synthesize large amounts of data using a range of programs. It is recommended that you work with data scientists/analysts/modelers who are proficient at geospatial analysis, as the geographic or locational component is key in hyperlocal air pollution mapping.



Air pollution experts or scientists

who provide expert advice in guiding analysis. These experts and scientists with intimate knowledge of air pollution can advise on factors critical to analyzing air pollution data, such as impact of meteorology, wind direction, air dispersion, etc. They are often scientists in the field of environmental engineering, atmospheric chemistry, atmospheric physics or air pollution transport.



Individuals with local knowledge of air quality sources and concerns

such as local community partners, community leaders and residents, or researchers with in-depth knowledge on the air quality concerns in a particular area. Local insights are extremely valuable in informing data interpretation.



STATIONARY MONITORING DATA ANALYSIS

For end users to make sense of the large amount of data, the team must aggregate and present individual measurements in meaningful ways. For instance, the Breathe London team aggregates one-minute NO₂ measurements to hourly average concentrations which appear on the Breathe London [air quality map](#).

Similarly, the 100x100 team aggregated black carbon sensor data from millions of 1-minute measurements to hourly average black carbon concentrations. This resulted in each of the 100 sampling sites having 2400 hours of data (24 hours x 100 days).⁹ The team compiled the dataset on a visualization [tool](#) that allows for the exploration and analysis of black carbon concentrations across different time periods.¹⁰ Data were analyzed to show hourly average black carbon levels at each monitoring site throughout the campaign period, or to see black carbon levels on a certain day of the week, or to compare levels on a weekday versus weekend. The results, embedded in the visualization tool, enable users to query the dataset to reveal different spatial and temporal patterns of air pollution.

MOBILE MONITORING DATA ANALYSIS

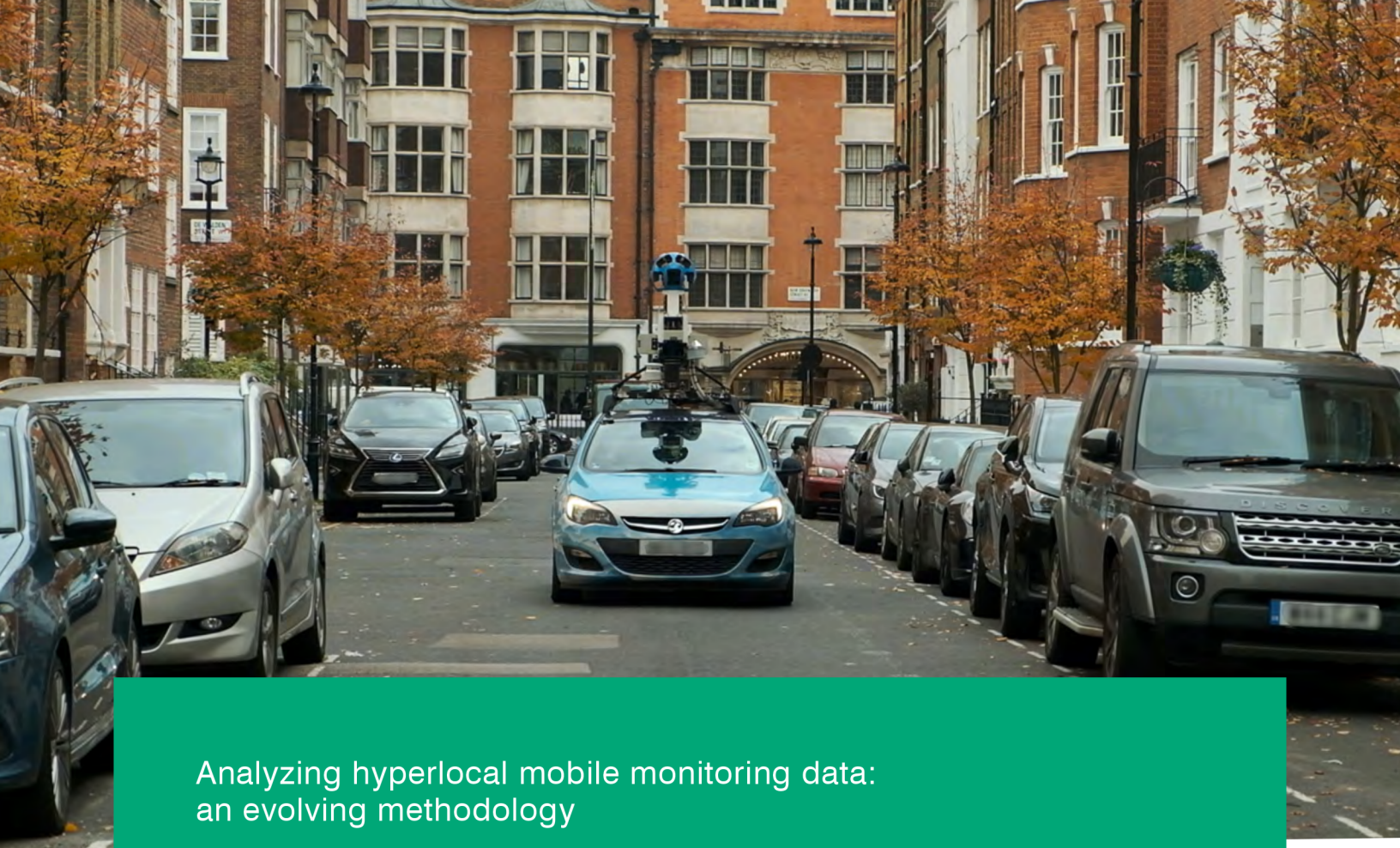
To render millions of data points collected from multiple months of mobile monitoring into a comprehensive air pollution map requires a deep and robust analysis. Below, we outline the basic steps that would allow you to conduct the sorts of analysis that EDF and partners have used in our hyperlocal mobile monitoring efforts. This analytical approach builds on a prior body of research.¹¹

The analysis can generally be broken down into four major parts:

- 1. Geolocation:** The process of assigning each air pollution measurement to uniform road segments (“grid-snapping”). The road segments should be of a uniform length to allow for an equal basis of comparison between segments. The length of road segments is based on what is appropriate for the speed of monitoring vehicles and instruments’ response time. Ultimately, the level of spatial resolution (that is, how granular a map you can generate) will depend on the response time of the instruments.
- 2. Assessing coverage:** Once you’ve assigned your air pollution measurements to the grids/road segments, you count them. You want to maximize the coverage, i.e. the number of drive passes over any road segment, that your budget allows. Each

measurement represents a snapshot in time, and you will want as many drive passes in different conditions (weather, background pollution levels, etc.) as possible, so you can get a representative average. The number of drive passes (or air pollution measurements) you end up with in your grid/road segments will determine the uncertainty level of your dataset. With more drive passes, the effects of anomalous events are smoothed out, resulting in a more representative dataset.

- 3. Finding the “central tendency”:** Involves statistical analysis to establish a measurement value that adequately represents the pollution concentration over a longer period of time (e.g. yearly or a multi-month monitoring campaign). Colloquially, you could think of it as a steady average or long-term average air pollution concentration at a particular grid/road segment. Options for representing central tendency include an average, a midpoint (a median), or the midpoint of many results that occurred close together. For example, in Oakland, mobile monitoring platforms collected about 200 unique observations for each 30-meter road segment. The pollution maps were generated by first calculating the average concentration within each 30-meter road segment during each drive pass, and then calculating the median of averages across multiple drive passes for that 30-meter segment. This algorithm reduces the influence of individual extreme samples if, for example, a monitoring car happened to be driving behind a truck during one pass of a street.¹²
- 4. Identifying hotspots:** Hotspots are the areas where the central tendency values are elevated compared to a reference value. Hotspots can be defined in multiple ways and will depend on the insights or actions for which the data is intended. For instance, you may want to identify the areas on the map where pollution concentrations tend to be above a certain threshold. You might compare the central tendency values (e.g. average black carbon concentration at a spot) with a chosen baseline concentration level. Generally this is done as a fraction, with the central tendency as the numerator, and a relevant baseline as the denominator, seeking those locations with fractions greater than one. Teams can use a neighborhood, a defined area, or a citywide average pollution level as the baseline. For example, in our initial Oakland work hotspots were defined as locations where concentrations of multiple pollutants exceed nearby ambient levels by 50% or more.¹³ In order for the spot to be considered elevated, the uncertainty around the central tendency value should not have substantial overlap with the uncertainty around the baseline value.



Analyzing hyperlocal mobile monitoring data: an evolving methodology

In mobile monitoring projects like the ones we've conducted in Oakland, Houston, and London, we generated a huge amount of data. For instance, in Oakland, the monitoring cars drove more than 14,000 miles of roads for more than 150 days, collecting data every second. This generated more than 3 million data points. As discussed above, we used data science to map each individual data point, representing one second of pollution observations at a specific location, to a corresponding 30-meter length of road. This algorithm allowed us to turn millions of data points into a map like the one shown below. The initial analytical method used in Oakland was published in a 2017 peer-reviewed paper.¹⁴ A summary of this methodology can be found on EDF's [website](#).

In Houston, our team updated the algorithm to calculate and more precisely reflect the actual distance a car travels between measurements, while previously we had used 30-meter distances in the Oakland project which was based on average car speed and instrument response time. This improved method better reflects the real world conditions under which data is collected. The methodology continues to evolve with our work in Breathe London to enable better characterization of the central tendency and hotspot identification.

While the analyses appears complex, cities can tackle them by working with the relevant experts and skilled technicians. EDF and others are increasingly making the algorithms and software necessary to make the work open-source and readily available. Algorithms used in our Houston and London work can be found at the following [Github repositories](#).

Beyond mapping air pollution patterns and identifying hotspots, we are exploring new ways to link hyperlocal data to a deeper understanding of emission sources and health effects that will allow individuals and policy makers to take additional actions to reduce adverse impacts.

We are developing methods to quantify and map the health risks and impacts of air pollution at a hyperlocal scale.

Sensor technology is making the variation in air pollution in cities visible. But populations and health risks are not uniformly distributed across urban areas. The risks and impacts of air pollution may not follow the same pattern as that of air pollutant concentrations alone, as factors like demographics and health status that impact sensitivity to air pollution also vary across cities.

By combining information on hyperlocal air pollution with knowledge of the impact that these pollutants have on health, population distribution, and variation in risk and disease susceptibility within a city, we can quantify and map the risks and impacts of air pollution within and across neighborhoods.

Going forward, we will incorporate information on source characterizations to develop estimates of contributions of different sources to the air pollution risks and impacts in communities. Decision makers can use this health impact analysis to better understand the places and populations most impacted to effectively target mitigation action. This will be ready in 2020.

CASE STUDY: DISTRIBUTION OF NO₂ POLLUTION AND ASSOCIATED MORTALITY RISKS AND IMPACTS IN NEIGHBORHOODS IN OAKLAND, CA

EDF has intensively measured and mapped air pollution in West and downtown Oakland. Nitrogen dioxide (NO₂), a traffic related pollutant, is known to cause several adverse health effects and increases the risk of premature death.

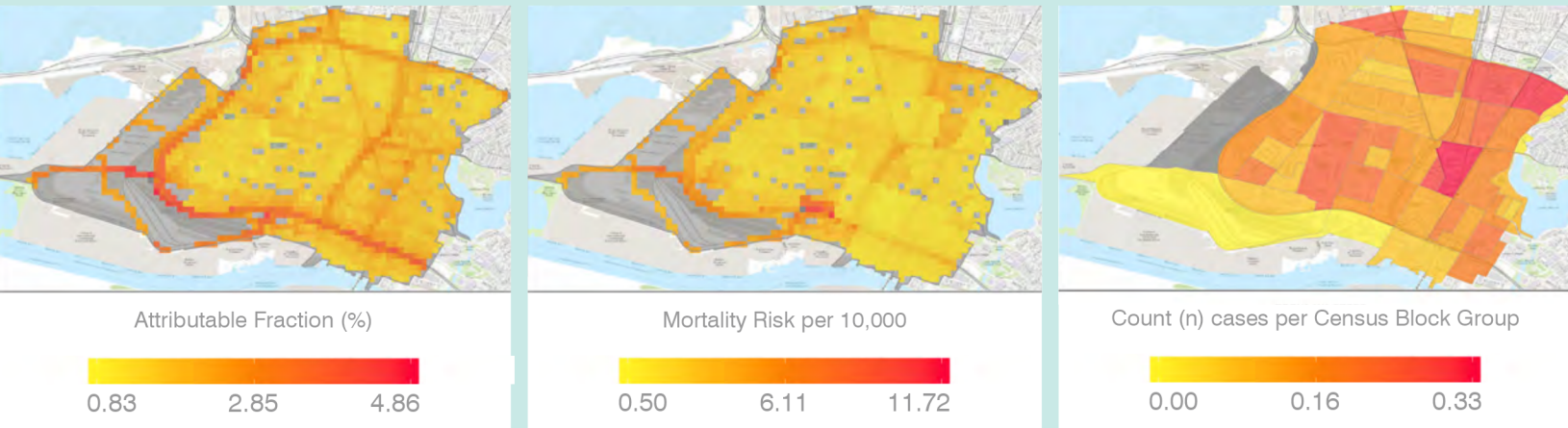
Preliminary results from our health impact study in this community indicate that NO₂ contributes to nearly 1 in 20 premature deaths (attributable fraction: 4.86%) in the areas with highest pollutant concentrations in comparison to <1 in 100 deaths (attributable fraction: 0.83%) in the areas with the lowest concentrations. However, baseline mortality rates vary from place to place due to other factors (such as poverty, race, nutrition, and health).

Using Alameda County Public Health Department mortality rates, at the census block group scale, we found that the distribution of the impact of NO₂, in terms of attributable *excess mortality risk*, was different than the pattern seen for pollutants alone, and disparities in risk were remarkable. The area with the most elevated risk (11.7 deaths per 10000 people), near the intersection of freeway I880 and I980, had twenty times higher risk than the lowest risk area (0.5 deaths per 10000 people). This is displayed below in Panel B.

Finally, incorporating the population and age data, and aggregating the results to census block group level, we were able to identify where the greatest adult (>25 yrs old) mortality impacts of NO₂ air pollution are experienced (Panel C.). This methodology reflects cumulative risks of air pollution on vulnerable populations. Without incorporating information on baseline mortality risks and population distribution, decision makers may miss important areas of at risk populations as well as the most impacted areas, where return on investments for mitigation of air pollution may be highest.

WEST OAKLAND: NO₂ ATTRIBUTABLE PREMATURE MORTALITY

Geographic area: 15 km²



A. Fraction of premature deaths attributable to NO₂ exposure

B. Excess mortality risk attributable to NO₂ exposure

C. Premature deaths per year attributable to NO₂ exposure

UNDERSTANDING SOURCE ATTRIBUTION

We are developing and testing methods that can be used to identify sources of measured air pollution using innovative, hyperlocal techniques. Our partners will use weather models to generate “footprints” describing the most likely path the air traveled on its way to being measured by the instruments. Local sources that contribute to elevated pollution levels are most likely located within these footprints. Because of uncertainties in wind speed, wind direction, and wind- and heat-driven atmospheric turbulence or mixing, these footprints resemble pieces of pie, expanding in width (and therefore uncertainty) as we move backwards in time from the measurement.

However, with many data points, measured in many locations on days with different weather patterns, we can narrow down modeled relationships between measured pollution and likely sources by overlaying these footprints to identify repeating patterns. In addition, we are also leading work to understand how improvements in meteorological models could make footprints more narrow, and possibly reduce the number of measurements needed to identify sources. We anticipate this analysis will be completed by mid-2020.



PART 03: DATA TO ACTION



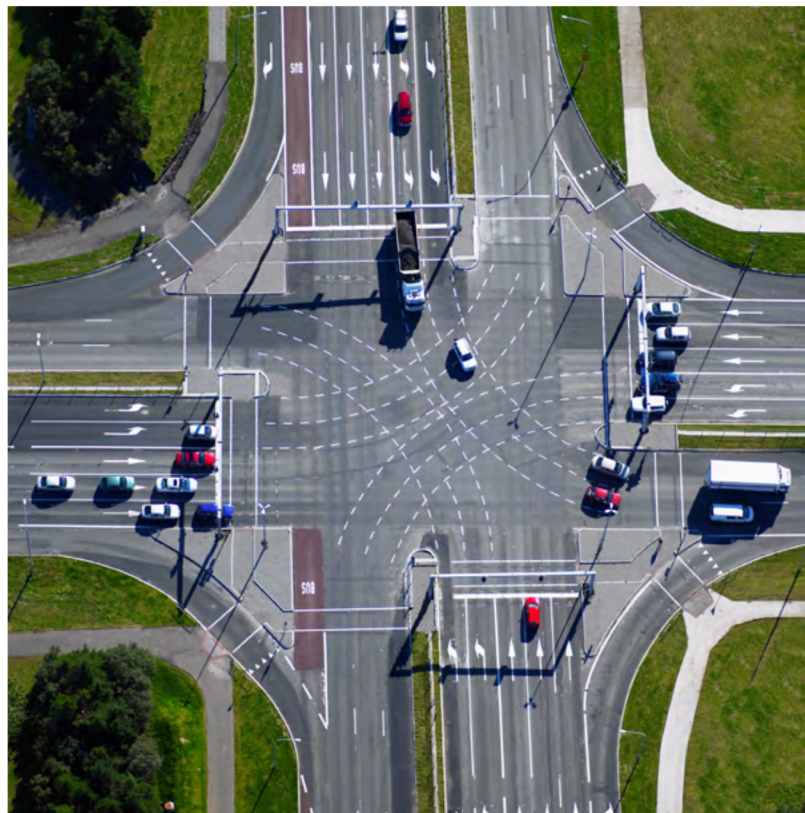
Using air pollution data to develop and implement clean air solutions

Once your team knows that the data is thorough and accurate, and once you visualize and interpret the results, you can present the facts and make the case for new policies. At their best, maps can stimulate engagement and direct people towards action. At their worst, content without context runs a real risk of numbing people to yet another threat they feel they can't do anything about. Anyone releasing hyperlocal maps should have a plan for continuing the conversation after the maps are released to explain the results and respond to questions from community members who want to use the data to enable decisions around possible steps to reduce pollution. This [resource](#) highlights many ways citizen scientists are using air pollution data to inform advocacy.

Once you've put a monitoring network in place and have ensured a reliable flow of quality data, remember that managing risk and working closely with stakeholders to outline policy changes involves many personalities and can take time. Impactful pollution reduction policies can take years to implement, long after you create your monitoring network and start working with communities. Therefore, securing funding and staffing that will sustain you for several years is necessary to maintain momentum toward significantly reducing air pollution.

The power of hyperlocal mapping will be felt in your community when a policy or action draws on the data you gathered. This can occur along many paths, each outlined below, and mapped to the possible policies and actions we listed in Part 1. Data can also support long-term changes in how your city works, and we'll close our guide with a tour through some of those changes.

- Investigation and enforcement against factories or other stationary pollution sources
- Emergency public health interventions, like evacuations, shelter in place, or public information campaigns
- Transportation planning (long-term)
- Traffic management (short-term)
- Zoning, permitting, building codes, and land use
- Targeting investments and incentives for emissions-reduction projects, like EV buses or building retrofits



Air quality dashboard

Many local leaders are already investigating and enforcing based on their existing authority, but believe they could be more efficient and effective. In Houston, we worked with Rice University to develop a dashboard so the city can target investigations better using data collected by municipal vehicles driving their regular routes. [Click here](#) for a case study on Houston's investigation dashboard.



Emergency public health interventions

If you want to coordinate work among public health officials after a natural disaster, review how we helped Houston identify elevated levels of benzene after Hurricane Harvey with this [case study](#). As a result, our team is working with Rice University and the City of Houston to build an online platform that will alert city officials of dangerous pollution levels early, before they become a public health concern.

TRANSPORTATION PLANNING AND TRAFFIC MANAGEMENT

Many people around the world increasingly consider a personal vehicle a hassle rather than an entitlement. They are willing to consider mobility as a service, rather than an asset. Investors bet [\\$40b in venture capital](#) between 2016 and 2018 on this shift. China has over 400,000 electric buses on the road. [Columbus, Ohio](#) won a \$50 million grant to offer to other cities its lessons in “detoxifying” transportation. Although emissions from transportation continue to grow, this too is changing. This shift is partly driven by city-led policies. Commercial vehicles in Europe will soon have to contend with zero-emissions urban access zones. Cities around the world are adopting [C40 Cities’ “fossil-fuel free street” goal](#). It aims to have cities procure zero-emission cars and remove polluting ones within a strategy of “people-friendly planning.”

Data-driven methods already exist for traffic and transportation planning:

- Google’s [Environmental Insights Explorer](#) combines inputs from Google maps with a tool called [CURB](#) to help estimate total tons of carbon produced per year in your city’s building and transportation emissions.
- The [Joaquin Decision Support Tool](#), which grew from an initiative in the European Community, helps decision makers and staff choose the best-fit measures to improve local air quality traffic policies. The tool organizes mitigation measures into several categories and scores each one based on a mix of literature reviews, case studies, and other inputs.

With these tools, you are able to explore improvements to baseline estimates by changing variables like total miles traveled by auto, bus, bike, foot, rail, subway, average vehicle efficiency, and emissions per type of vehicle.

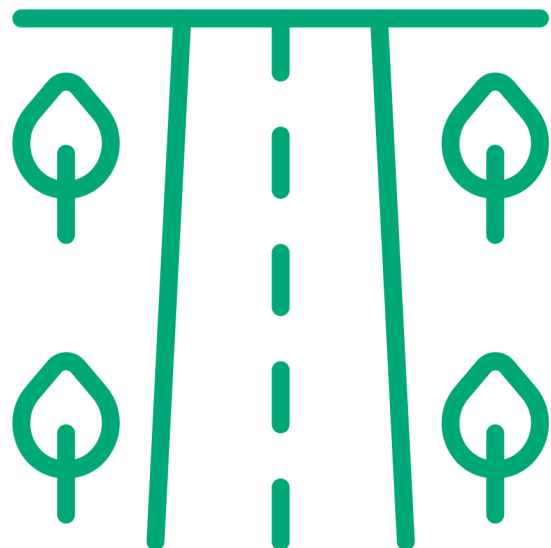
These tools, and peer-reviewed health studies and guidelines, while founded in sound science, lack the hyperlocal data clean air advocates often need to advance new policies in their cities. Hyperlocal insights offer the opportunity to pinpoint the risks and dangers of pollution where people live, work, and play. This gives advocates a rallying point to build a base of support, and gives policymakers the ability to develop targeted traffic and transportation interventions to maximize the benefits and minimize the cost of action. For example, this [case study](#) shows how community advocates used hyperlocal data from our project in Oakland to influence local traffic planning decision.

ZONING, PERMITTING, BUILDING CODES, AND LAND USE

By spotlighting both immediate and long-term perils from pollution, city governments can build a case for changes in land use or zoning. Some innovations in recent years include:

- **Pedestrian zones or car-free zones.** With these, officials designate areas of a city or town for people on foot or on bikes, and may ban all vehicles. Some experts [contend](#) these policies guide transportation officials away from prioritizing traffic and toward celebrating people. Other advocates seek similar shifts by urging cities to [end free parking](#).
- **Clean construction requirements.** Hyperlocal data that shows hotspots or elevations at particular times can bolster support for [clean construction initiatives](#). This approach, which has driven [legislation in Pittsburgh](#), involves requiring low emissions and mitigation equipment on all new construction projects.
- **Zoning and permitting in environmental justice communities.** [Low-income communities](#) and communities of color experience a disproportionate burden from land use patterns and the associated impacts of polluting industries. With hyperlocal data, you can design long-term changes that can improve the imbalance. This [national scan compiles 40 policies](#) from more than 20 cities across the U.S. It includes six types of policies, from bans on certain kinds of polluting facilities to new environmental review processes to proactive planning and new health codes.

All of these policies could be more attractive, and more effective, when targeted with hyperlocal air insights.



TARGET INVESTMENTS AND INCENTIVES

As public authorities explore the value of hyperlocal air pollution data, they can work with communities to use the data to steer funds toward efforts to improve air quality. For example, [AB617](#) in California welcomes highly granular air pollution data. Legislators drafted the bill to direct air agencies to consider localized data when investing in community air pollution mitigation efforts. With leaders paying smarter attention to environmental concerns around the world, hyperlocal air insights could increase the efficacy of public investments by ensuring the funds go to neighborhoods with the greatest air pollution burden.

Investment programs that could create greater impact by using hyperlocal air pollution insights include:

- **Vehicle replacement funds:** In the U.S., Diesel Emissions Reduction Act funds and Volkswagen Settlement funds currently help pay for transportation improvements. Cities could provide data to assist in project selection.
- **City-owned assets:** Cities are owners and managers of vehicle fleets and buildings. Investments in improvements such as EVs can be targeted to air pollution hotspots first.
- **Private assets:** The private sector is constantly investing in new vehicles, building stock, and other potential sources of pollution. Many private sector leaders have committed to climate goals, but have yet to create implementation plans. Cities could provide data and a convening pressure to encourage the private sector to target climate-driven investments to provide the greatest two-for-one health and climate benefits.
- **Building incentives:** These can include incentives to replace wood stoves or furnaces, and energy retrofit or energy efficiency funds in cities or across broader areas.
- **Green banks and public-private impact funds:** Investors and investment advisors can use hyperlocal data to prioritize projects in air-stressed neighborhoods.
- **Energy efficiency for industrial combustion:** The World Bank manages loan pools for cleaner factories or other potential polluters. Cities can provide hyperlocal information to prioritize certain areas.
- **Tax rebates and faster depreciation:** These can be directed to monitors and related equipment.



As public authorities explore the value of hyperlocal air pollution data, they can work with communities to use the data to steer funds toward efforts to improve air quality.