Trash Data Dive Projects

# Comprehensive Trash Stories (Total Trash Picture)

## Challenge / Problem Statement

Integrate disparate trash-related datasets covering a defined geographic area to describe and derive insights into the causes and effects of trash pollution in that area.

## Background

Data directly related to trash composition, abundance, or distribution (i.e. ‘primary’ trash data) often comes in a variety of forms. For example, primary datasets may be in completely different formats (e.g., quantitative tabular data, qualitative assessments, or image libraries), measure different attributes (e.g. volume, type, or material), assess different parts of the trash lifecycle (e.g. material collected by waste management systems for disposal or recycling, or material found outside waste management systems as litter), or measure at different geographic or temporal scales (e.g. trash found at individual control devices or total trash collected from a given area, for individual events or at regular intervals across time). As a result, different trash datasets covering the same geographic area are often treated in isolation.

In addition, there are a virtually unlimited number of ‘secondary’ datasets – such as socio-economic/demographic, economic/commercial, land use, climate/precipitation, environmental/ecological – which may be combined with or compared to primary trash data to uncover trends and correlations. These investigations could help to further describe the causes and effects of trash pollution, and could also help to extend insights and conclusions derived from data with a limited scope across a broader range.

## Questions

* Within a given geographic area, how can a variety of different primary and secondary datasets be combined to provide a comprehensive assessment of trash pollution (composition, abundance, distribution, etc.)?
* What types of relationships between primary and secondary datasets can be observed to make inferences about the causes and effects of trash pollution, and what are the implications of those relationships for various stakeholders (managers, regulators, industry, citizens, etc)?
* Can the conclusions from this assessment be extended to other geographic areas?
* Are there gaps in the data? Are there particular types of data that could be collected in this area that would be particularly useful?
* Can any patterns of observation be leveraged to help predict trash generation rates? Do we see temporal or geospatial patterns?

## Skillsets Needed

* Data cleaning and formatting
* Data integration (and data schemas), visualization, and mapping
* Statistics (correlation, trend analysis, prediction, etc.)
* Knowledge and understanding of policy and context (e.g., need to understand what questions policymakers and regulators are asking in order to provide useful answers and actionable information, need to investigate correlated factors to identify ultimate causes, etc.)

## Resources

Provided Resources:

* Data from existing monitoring programs (e.g. City of Sacramento’s On Land Visual Assessments, BASMAA’s On Land Visual Assessments)
* Relevant image libraries from machine learning projects

Other Potential Resources:

* Demographic / socioeconomic data (e.g. population distribution, age, income)
* Land use and economic data (e.g. location of particular stores, commercial centers, or public gathering places)
* GIS data for waterways and stormwater infrastructure

# Machine Learning Frameworks for Trash Monitoring and Compliance

## Challenge / Problem Statement

Evaluate and refine existing frameworks that apply computer vision for trash monitoring (or develop new ones) to make them more applicable to specific regulatory and management questions.

## Background

Across California, several independent efforts are underway to investigate ways to improve the effectiveness and efficiency of trash monitoring programs by applying large-scale image collection programs and machine learning. In general, the idea behind these projects is to take advantage of recent developments in technologies that make it relatively easy and inexpensive to capture and store large volumes of images in areas where trash is a potential concern (such as sensitive areas like waterways, or areas where trash could easily be mobilized to enter sensitive areas like roadways and other areas connected to stormwater infrastructure). Computer vision models (a subset of machine learning) can be trained to detect and describe various aspects of trash using a subset of these images, and once optimized those models can be used to process newly acquired imagery. Monitoring programs can then be implemented which use large-scale photo or video monitoring to provide extensive coverage of sensitive areas across space and time, and the customized computer vision models can be applied to process the newly acquired imagery in an efficient and objective (i.e. standardized and reproducible) manner, in a way that would not be possible with manual processing techniques. Some examples include efforts undertaken by CalTrans, Litterati, a partnership between SCCWRP and SFEI, and a partnership between the California State Water Resources Control Board and the City of West Sacramento.

However, there are also important differences in the scope of these projects, including the types of image data collected and the features assessed by the computer vision models. The various projects collect images from cameras affixed to vehicles that travel throughout a study area (CalTrans and Water Board), drones (SCCWRP / SFEI), or crowdsourced efforts from individuals (Litterati). As a result, the images may show trash on vastly different scales, with variable levels of detail and in very different contexts. In addition, the models differ in the features of trash pollution that they attempt to recognize and describe. For example, some models may simply assess the presence or absence of trash, while others variously attempt to evaluate the total volume or count of trash objects within an image, or describe features like type, material, or brand of individual trash items. Coordinating certain aspects of the models associated with each project could help to provide more value by increasing the interoperability of their inputs and/or outputs.

While these efforts all provide valuable information in their own right, the information they currently provide may not necessarily meet the immediate needs of regulators and managers, who often require specific information to make decisions and meet regulatory requirements. If this incongruity is not addressed, regulators and managers may be slow to adopt of these new techniques, and the impact of these approaches may be greatly diminished. For example, the California Water Resources Control Board currently accepts two pathways for compliance with permits that govern regulated entities that are potential sources of trash pollution in waterways. One of those methods relies on a qualitative visual assessment of sensitive areas to assess trash pollution (referred to as the On Land Visual Assessment approach). If current machine learning methods could be adapted to satisfy the demands of this framework and reliably provide assessments similar to those performed by humans, then regulators could likely accept this approach as proof of compliance with regulatory requirements. This could significantly increase the rate of adoption of this approach, and make its benefits (e.g. providing more actionable information, making analyses more standardized and reproducible, and reducing costs) much more widely available.

## Questions

* What steps are needed to apply existing computer vision models and image libraries to develop assessments consistent with the On Land Visual Assessment regulatory framework currently used by the State Water Board to assess permit compliance?
* Are there other computer vision models / frameworks that could be applied to improve on the effectiveness or efficiency of existing models?
* What attributes of trash should be collected and used to train the computer vision models (e.g., classification, material, brand, count, volume)?
* Are there ways to translate existing datasets used for certain models to make them interoperable with other models?
* What schema should future data collection efforts use to optimize usability and interoperability?
* How can machine learning/computer vision be used to increase the data quality of citizen science efforts?

## Skillsets Needed

* Machine learning / computer vision
* Image processing
* General data analysis, statistics, and geospatial analysis (to aggregate results for a particular place/time and translate into a single score)
* Knowledge of existing regulatory frameworks and compliance pathways related to trash (specifically the On Land Visual Assessment methodology)

## Resources

Provided Resources:

* Image libraries (along with annotations and metadata input to existing machine learning models)
* Model code (for projects where it’s publicly available – likely CalTrans, Waterboard, SCCWRP/SFEI)
* Documents describing the On Land Visual Assessment methodology (and reports describing any efforts to apply existing computer vision frameworks to this methodology – e.g. CalTrans)

Other Potential Resources:

* Public image libraries
* Open-source computer vision models and image processing tools

# Tobacco-Related Trash Assessment and Management

## Challenge / Problem Statement

Integrate and analyze datasets related to tobacco products and their contribution to trash pollution in California to find ways to better describe and manage the problem of tobacco-related pollution in waterways.

## Background

The significant contribution of tobacco products to trash pollution in California’s waterways has been well documented, and studies have shown the likely impact of that pollution on environmental and public health[[1]](#footnote-0). As a result, the subset of trash pollution related to tobacco products is a high priority for regulators, managers, and the public. However, efforts to reduce tobacco-related pollution have been largely unsuccessful, and new approaches are needed. Using data to better describe the problem and develop targeted solutions could be a way to make progress on this problem.

## Questions

* What is the magnitude of tobacco-related trash pollution? How does this compare with other types of trash pollution?
* How is tobacco-related trash pollution distributed geographically and temporally?
* What are the characteristics of tobacco related pollution (e.g. type, brand), what is the relative magnitude of tobacco-related pollution in those categories, and how does it vary geographically and temporally?
* What secondary factors influence the magnitude and characteristics of tobacco-related trash pollution, and how might those insights be applied to manage and reduce the magnitude of the problem?
* Can we correlate high tobacco loads in trash to adverse environmental impacts, such as wildlife, vegetation, and drinking water impairments? What is the fate of tobacco products that enter the marine ecosystem?

## Skillsets Needed

* Data cleaning and formatting
* Data integration (and data schemas), data visualization, and mapping
* Statistics (correlation, trend analysis, prediction, etc.)
* Knowledge and understanding of policy and context

## Resources

Provided Resources:

* Tobacco-related survey data (from CDPH)
* Image libraries from computer vision projects
* Southern California Stormwater Monitoring Coalition Stream/River monitoring data

Other Potential Resources:

* Tobacco sales locations
* Tobacco sales data
* Demographic and socioeconomic data
* Environmental and public health data
* Impacts: CECs (Chemicals of Emerging Concern) and microplastics/microfiber data

1. https://www.newsdeeply.com/water/articles/2018/10/17/is-resistance-futile-cigarette-butts-still-dominate-coastal-litter [↑](#footnote-ref-0)