Washington State Department of Ecology

Data Handling Report #1: Methodology, Universe Identification, and Sample Design – Auto Body Pilot

This document explains the preliminary methodology and data screening techniques used for the Sustainable Washington Autobody pilot.

Universe Definition

The pilot program targets all autobody repair shops in each of the local jurisdictions receiving a Local Source Control grant from Ecology. These jurisdictions are:

City of Bellevue
City of Bellingham
City of Issaquah
King County
Kitsap County
Pierce County
San Juan County
Skagit County
Snohomish County
Spokane County
Whatcom County

There is some overlap between some jurisdictions. The cities of Bellevue and Issaquah are both in King County, and the city of Bellingham is in Whatcom County. In these cases, we allocated shops based on the smallest applicable legal jurisdiction. That is, shops whose physical location fell within the city limits of Bellevue, Bellingham, or Issaquah were allocated to those cities; shops that did not fall within the city limits were allocated to the county, even if the mailing address was Bellevue, Bellingham, or Issaquah.

Determinations of city and county were made by checking each address in the Washington State Department of Revenue's taxing district GIS system, which is available at http://www.dor.wa.gov/content/findtaxesandrates/salesandusetaxrates/lookupataxrate/.

Universe Identification

Potential facilities were identified using three sources: Harris InfoSource, the Washington Department of Labor and Industries, and applicable local air authorities.

Harris InfoSource is a Dun & Bradstreet company which markets a database called "Selectory." We used the Selectory database to produce a list of all companies in the target geographic area with a NAICS code of 811121 ("Automotive Body, Paint, and Interior Repair and Maintenance"). This NAICS code classification also includes shops that perform auto and marine upholstery work and other custom auto interior work. Although not autobody repair shops, we left these entries on the list because establishments of this type may also provide similar services to a repair shop. For example, a shop that only provides pin striping services would also be in this NAICS code; while they do not repair cars, they do apply auto paint and will share many issues that our program is trying to address. Using the Selectory database produced a list of 961 facilities.

The Washington State Department of Labor and Industries ("L&I") provided a list of all facilities in the state that have employees working in classification 3412-00 ("Auto or Truck Body Shop"). This industry classification also includes shops that apply spray-on bed liners for pickup trucks. Although not autobody repair shops, we left these entries on the list because we did not have a reliable way to distinguish shops only applying bed liners from those performing repair services. In addition, we believe shops applying bed liners will have many of the same issues as autobody repair shops, especially related to hazardous chemicals, air pollution, and water quality. Using L&I's database produced a list of 764 facilities.

We also checked the Puget Sound Clean Air Agency, Northwest Clean Air Agency, and Spokane Regional Clean Air Agency to verify whether any autobody shops in their jurisdictions had an air permit. No facilities were found that had not already been identified from either the Harris or L&I lists.

Initial Data "Cleaning"

The lists of potential program participants were "cleaned" to remove extraneous facilities. We reviewed and deleted the following from the list of facilities:

- Facilities not located in one of the target jurisdictions (e.g., Yakima County)
- Facilities clearly not autobody repair shops (e.g., an attorney's office)
- Facilities already on the list (duplicate entries)

This round of cleaning produced a list of 947 facilities.

After using the Harris InfoSource database for a similar project, we learned this database contained up to 75% incorrect entries (primarily of companies that had gone out of business or moved location). As a result, we decided to clean the list again. Each entry on the list was compared to the business records databases from the Washington State

Department of Revenue, Washington State Department of Licensing, and the Washington Secretary of State Corporations Division.

> If a business with the same or substantially similar name and address was found in at least one state database, the business was presumed to be active and was left on the list.
>
> If a business with the same or substantially similar name was found with a second address, the business was presumed to be active and was left on the list at the original address. This was because the second address may be a mailing address or an owner's home address. If it is a former address and is selected for a site visit, an outdated address will be updated for the second round of site visits.
>
> If a business with the same or substantially similar name was not found in any database, the business was left on the list. While it is possible these "can't find" entries are either out of business or have moved, it is also possible that they are open and active at the listed address but have a different legal name (e.g., "Joe's Autobody Shop" on our list might actually be a trade name for the legal entity "Smith and Sons, Inc." – the name that would be in the state databases.)
>
> If a business with the same or substantially similar name was found but listed as "closed" in either the Department of Revenue or Department of Licensing databases, the business was presumed to no longer be in business and was removed from the list. If a "closed" entry was only found in the Secretary of State's database, it was not removed. This is because the Departments of Revenue and Licensing require an action to close an account. The Secretary of State will close an account if it is not properly renewed each year. So if a Revenue account is listed as "closed," an owner took an action to close the account. But if the Secretary of State reports a company is "inactive," it may be attributable to a paperwork oversight and the company may actually still be in business.

After this additional round of cleaning, the list was reduced to 831 potential participants. This reduction—just over 12% of facilities—is far less than the 75% we expected to find. We attribute this difference to our source data. In addition to the Harris InfoSource data, we also used the database from L&I. The L&I database is a list of businesses that have affirmatively represented to the state that they are open for business and have employees working in the autobody repair industry. More importantly, these businesses are paying workers' compensation rates based on this representation. If a business closes or no longer employs workers performing dangerous (and expensive) work such as autobody repair, it is in their best interest to report this change to L&I as soon as possible. Therefore, we believe this database is probably the most accurate one available to us and this high level of accuracy improved the overall accuracy of our list.

However, it is important to note that we did not elect to use the L&I database exclusively because it has the potential of being too restrictive. There is a significant possibility that there are autobody repair shops operating in the pilot program's target area that do not

appear on L&I's list. Autobody repair is a dangerous line of work. Businesses in this industry are required to pay relatively high workers' compensation rates. Thus, there is an incentive for a company to misrepresent the type of work they do when making their reports to L&I. If they are not caught, an autobody shop could potentially save money on its workers' compensation premiums by claiming its employees perform "auto repair" instead of "autobody repair." In order to cast the widest net possible and ensure that we maximize our potential for finding these shops, we used the Selectory database (which is less accurate but is more likely to find these shops) in addition to the L&I database.

Original Sample Design

Our initial sample design was to identify a statistically significant portion of our universe. We planned to use an Excel spreadsheet available from USEPA that performs the calculations based on universe size, confidence level, and margin of error. The universe of facilities would be stratified based on local jurisdiction and the calculated sample size would be allocated proportionally among the various local jurisdictions. Each jurisdiction would be a separate stratum and would receive a list of facilities to visit. The lists would be prepared by Ecology and would be randomized using a random number list generated by www.random.org.

Sample Size Calculation

We used EPA's "ERP Sample Planner" to calculate our sample size, using the formula based on a specific margin of error for a two-sample test. The planner uses the following formula[1] to calculate a sample size based on a given margin of error and population:

$$n = \frac{(Z_\alpha)^2[P_1(1-P_1) + P_2(1-P_2)]}{\delta^2}$$

Based on this formula, for a universe of 831 facilities, the planner calculates a sample size of 153 site visits for each of two rounds of inspections (given a 90% confidence level and a margin of error ±8.5%). This is equal to visiting 18.4% of the total universe in each of two rounds of site visits.

Stratification and Sample Selection

The list of 831 facilities was divided into strata based on the jurisdiction of the physical address listed. As noted above, facilities located in more than one relevant jurisdiction were assigned to the strata for their relevant city. When stratified, the universe was divided as follows:

City of Bellevue	24
City of Bellingham	13
City of Issaquah	9
King County	258

[1] Derived from Kish, Leslie. (1965). *Survey Sampling.* New York: John Wiley & Sons, Inc.

Kitsap County	92
Pierce County	126
San Juan County	4
Skagit County	33
Snohomish County	132
Spokane County	101
Whatcom County	39

However, when we reviewed the stratified lists more carefully, we discovered facilities that were incorrectly listed. Some facilities were listed in the wrong county – such as a facility in Everett (Snohomish County) being listed in Pierce County. We further cleaned the list, ensuring each facility's jurisdiction was properly identified and removing any facilities that were incorrectly listed as being in one of the target jurisdictions. One facility was on the wrong list because of an error with the street address; we corrected the address and placed it on the proper list. This review and revision resulted in a final universe list of 779 facilities.

The final universe of 779 facilities was stratified as follows:

City of Bellevue	23
City of Bellingham	13
City of Issaquah	9
King County	241
Kitsap County	77
Pierce County	124
San Juan County	4
Skagit County	32
Snohomish County	128
Spokane County	99
Whatcom County	29

Using the final universe numbers, we recalculated the number of site visits required for a statistically valid sample at 151. The proportion of facilities receiving site visits is therefore 19.4%.

We then multiplied each stratum's total number of facilities by 19.4%, resulting in each jurisdiction performing the following number of inspections:

City of Bellevue	5
City of Bellingham	3
City of Issaquah	2
King County	47
Kitsap County	15
Pierce County	25
San Juan County	1
Skagit County	7
Snohomish County	25
Spokane County	20

Whatcom County	6

These numbers are identical to the alternative method for sample calculation. If we use the formula for determining sample size for proportional allocation in a stratified sample

$$n_i = \frac{N_i}{N} \cdot n \quad for\ i = 1, 2, \ldots, and\ k.$$

In order to guarantee an adequate number of site visits, we have rounded each stratum's numbers up to the next whole integer in both methods explained above.[2] This resulted in 156 total site visits to be performed by the local jurisdictions.

Small Strata Methodology

The above methodology creates five strata with a sample size of greater than 15 sites per strata. We believe this is adequate and will not negatively affect statistical calculations. However, the other six strata have sample sizes of less than 15 sites. These small strata have the potential to affect calculations of variance (and, consequently, any calculation incorporating variance). Fortunately, only 14.68% of the total number of facilities fall into these strata, so any potential effect on variance calculations should be very small.

Despite the small chance of affecting the outcome, we decided to compensate for these small strata by performing additional site visits. For the City of Bellingham, City of Issaquah, and San Juan County, we decided to use a census approach instead of a random sample. We chose these three jurisdictions because they each had fewer than 15 facilities in the entire stratum. While the City of Bellevue, Skagit County, and Whatcom County each also comprised a small stratum, the number of facilities in each of these jurisdictions was noticeably more than the smaller three strata. For these slightly larger (but still small) strata, we decided to oversample each stratum by 5%.

As a result, we divided the strata into three distinct groups:

1. *Jurisdictions with more than 15 total facilities and more than 15 site visits.*
This group will perform their proportional share of site visits as calculated above. Strata falling into this group are King County, Kitsap County, Pierce County, Snohomish County, and Spokane County.

2. *Jurisdictions with more than 15 total facilities but fewer than 15 site visits.*
This group will perform an oversample of site visits to compensate for the small sample size. Instead of sampling 19.4% of facilities in their jurisdiction, these jurisdictions will sample at least 24.4% (5% more, rounded up to the next whole integer) of the facilities in this jurisdiction. This translates to 1-2 extra site visits per jurisdiction. Strata falling into this group are the City of Bellevue, Skagit County, and Whatcom County.

[2] This is the case even when normal rounding rules would result in a number being rounded down. For example, a jurisdiction with a proportional share of 10.025 site visits would be asked to perform 11 visits.

3. *Jurisdictions with fewer than 15 total facilities.* This group will not sample their facilities, but instead will perform a census of all facilities in the jurisdiction. If necessary, Ecology staff will assist local jurisdiction staff in performing the site visits. Strata falling into this group are the City of Bellingham, City of Issaquah, and San Juan County.

Therefore, the final inspection numbers for each jurisdiction are as follows:

City of Bellevue	6	(1 extra visit)
City of Bellingham	13	(10 extra visits)
City of Issaquah	9	(7 extra visits)
King County	47	
Kitsap County	15	
Pierce County	25	
San Juan County	4	(3 extra visits)
Skagit County	8	(1 extra visit)
Snohomish County	25	
Spokane County	20	
Whatcom County	8	(2 extra visits)

When performing any statistical analysis on the data collected, we will use a weighted approach for the data from groups 2 and 3. Data collected from these groups will be weighted so that each group of strata is represented proportionally. When appropriate, data from group 2 will be multiplied by a factor of 0.8181 and data from group 3 will be multiplied by a factor of 0.2308 to ensure each group is not disproportionally represented.

Site Visit List Generation

After finally determining the applicable sample size for each stratum, Ecology generated a custom site visit list for each jurisdiction. Each stratum's list of facilities was organized in an Excel spreadsheet in alphabetical order. The line number corresponding to each facility's entry in the spreadsheet designated that unique facility. The number of total facilities for each stratum was then entered into the "Random Sequence Generator" available at www.random.org. The generator then produced a randomized sequence of the total number of facilities. We then used the sequence to create a customized random list for each jurisdiction's site visits.

For example, say the applicable stratum had 100 facilities and the generator produced a sequence that started with "46, 72, 9, 84, 28." The first facility to receive a site visit would be the facility listed on line 46 of the Excel spreadsheet, followed by the facility on line 72, then the facility on line 9, and so on.