Workplan to Identify Who is Underserved in the SMB Sector

[1 Outstanding Issues 2](#_Toc49334840)

[1.1 Initial Questions 2](#_Toc49334841)

[1.2 Areas of concern 2](#_Toc49334842)

[2 Methodological Overview 2](#_Toc49334843)

[3 Exploration of Participation Data 3](#_Toc49334844)

[3.1 Data Exploration Questions 3](#_Toc49334845)

[3.2 Data Source 3](#_Toc49334846)

[3.3 Definition of SMB 3](#_Toc49334847)

[3.4 Variables to Explore for describing SMB Participation 3](#_Toc49334848)

[3.4.1 Directly Observed Variables 3](#_Toc49334849)

[3.4.2 Synthetic Variables 4](#_Toc49334850)

[4 Spatial Analysis for Gaps by Geographic Area 4](#_Toc49334851)

[4.1 Data Exploration Questions 4](#_Toc49334852)

[4.2 Data Source and Synthesis 4](#_Toc49334853)

[4.3 Gaps in Participation by Geographic Area 5](#_Toc49334854)

[4.4 Gaps in Investment by Geographic Area 5](#_Toc49334855)

[4.5 Gaps in Savings by Geographic Area 6](#_Toc49334856)

[5 Spatial Analysis for Gaps by Socio-demographics and Geographic Area 6](#_Toc49334857)

[5.1 Data Exploration Questions 6](#_Toc49334858)

[5.2 Data Source 6](#_Toc49334859)

[5.3 Gaps in Participation by Socio-Demographic Groups and Geographic Area 7](#_Toc49334860)

# Outstanding Issues

## Initial Questions

1. What is the difference between the “CEI\_metadata” and the Claim\_metadata”?
2. How is SMB defined? Is it SMB defined by the “Claim” table, field “Sector” which is described as “Measure-level sector conforming to the Business Plan Sectors listed in CEDARS”? Is it based on rate class?
3. Do we have any access to usage data?
4. Do we have access to total customer or SMB counts by zip-code?

## Areas of concern

1. Defining SMB – We need a clear definition for SMB, see the question above. Otherwise we might be able to classify based on RateClass
2. Participation in upstream/midstream programs

# Methodological Overview

# Exploration of Participation Data

## Data Exploration Questions

1. What does program participation by SMBs look like?
	1. Review of summary statistics for continuous variables
	2. Frequency of categorical data
	3. Comparison with non-SMB pa
2. What correlates with or predicts SMB participation?
	1. T-test
	2. Multivariate regression analysis
	3. Other?

## Data Source

The primary energy efficiency activity data source should be the “Claim” table from CEDARS. This contains comprehensive information for individual rebates, including:

* Savings
* Costs
* Dates
* Rates
* Basic building info
* Site information

The Claims table does not appear to be for upstream/midstream programs. This is important to understand. For ease of analysis, we may want to exclude these types of programs, but ideally would include them.

The “ProgramCost” table may be helpful for additional analysis on marketing spending. It is not assigned to any location, but may be linked through program and year.

One area that would be nice to get some additional data would be on energy usage overall, and not just participation in programs

## Definition of SMB

We need to agree on a specific definition that aligns with the CPUC. Is it a business category or a rate class? Do we also want to include a more broad or narrow definition?

## Variables to Explore for describing SMB Participation

### Directly Observed Variables

A list of variables that make sense to look at as part of the statistical analysis are included as ***Appendix A – Data Sources***. This is based on previous experience with energy efficiency data exploration, and is a list of the fields in the main data sources with the variables that may be worth examining are indicated.

Given the large number of fields available, feature selection algorithms could be used to examine additional explanatory variables (random forest, ANOVA, etc.)

### Synthetic Variables

Based on previous experiencing analyzing energy efficiency data, I recommend that the following synthetic variable are also worth including in the statistical analysis:

1. % of Cost Covered = incentive amount / measure cost
2. Percentage Savings or depth of savings = savings divided by usage
3. Incentive / kWh
4. Incentive / therm
5. Measure cost / kWh
6. Measure cost / therm

# Spatial Analysis for Gaps by Geographic Area

## Data Exploration Questions

1. What does the spatial distribution of a variable look like?
	1. Identification of areas with high values (top quartile, top 10%, etc.)
	2. Identification of areas with low values (bottom quartile, bottom 10%, etc.)
	3. Heatmaps
	4. Clustering analysis
		1. Visual examination of histograms
		2. K-means
		3. KNN
2. How do variables change over time?
	1. 1 or 2 year period compared to previous periods

## Data Source and Synthesis

The Claims table includes a field for SiteID which references the following “Site” table in CEDARS.

|  |  |  |
| --- | --- | --- |
| **TableName** | **FieldName** | **FieldDescription** |
| Site | NAICSCode | The site NAICS code (eg. 112910) |
| Site | Residential\_Flag | Flag to identify Residential sites |
| Site | SiteCity | The site city |
| Site | SiteID | Unique and persistent identifier of each Site record |
| Site | SiteState | The site state, expected to be "CA" |
| Site | SiteZipCode | The site zip code, a valid California zip code |

The most granular value for site is SiteZipCode, which will probably limit our spatial analysis to the zip-code level, unless we want to use some sort of mapping between Zip/NAICS code.

The variables described in the following sections should be calculated over the time period and at the spatial level specified.

## Gaps in Participation by Geographic Area

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research****Question** | **Indicator** | **Measurement** | **Data Sources** | **Data****Analysis** |
| Are there gaps in program participation by geographic areas? | Participation | Number of businesses participating by neighborhood | PAs (Cedars)The CensusBureau ACS5-estimatesTiger/ Line Shapefiles | GIS SpatialAnalysis informed by statistical analysis |

**Spatial Granularity:** Zip-code

**Time Frame**: Annual. The count of all SMBs has to be done as snapshot in time, and so will not reflect changing business trends as new firms are started and others go out of business. The longer the look back period, the muddier this gets. Probably only possible to get this for the most recent year.

**Variable(s) to Explore:**

1. *SMBpart = Count of unique SMBs with at least one claim*
2. *SMBunique = Number of unique SMBs*
3. *SMB Penetration Level = Count of SMBpart / Count of SMBunique*
4. *NonSMBpart = Count of unique SiteIDs that are NOT SMB*
5. *Totalpart = SMBpart + NonSMBpart*
6. *SMB%TOTAL = SMPpart / TotalPart*

## Gaps in Investment by Geographic Area

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research****Question** | **Indicator** | **Measurement** | **Data Sources** | **Data****Analysis** |
| Are there gaps in program participation by geographic areas? |  Investment |  Dollars |  PAs (Cedars) | GIS SpatialAnalysis informed by statistical analysis |

**Spatial Granularity:** Zip-code

**Time Frame**: Annual, up to as many years back as possible.

**Variable(s) to Explore:**

1. *SMBincent = Sum of incentives if claim is for SMB*
2. *$/SMB = SMBincent / Count of unique SMBs with claims*
3. *%forSMB = SMBincent / Sum of all incentives*

## Gaps in Savings by Geographic Area

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research****Question** | **Indicator** | **Measurement** | **Data Sources** | **Data****Analysis** |
| Are there gaps in program participation by geographic areas? |  Energy Savings |  kWh/MW/therms saved per business |  PAs (Cedars), CEC Energy | GIS SpatialAnalysis informed by statistical analysis |

**Spatial Granularity:** Zip-code

**Time Frame**: Annual, up to as many years back as possible.

**Variable(s) to Explore:**

The following variables are defined for kWh, but work just as well for kW and therm.

1. *kWh/SMB = total kWh Savings (gross) for SMB / SMBpart*
2. *kWh%forSMB = Sum of kWh (gross) for SMB / Sum of all kWh (gross)*

## Data Exploration Questions

1. What other variables correlate with spatial areas with low/high SMB indicators identified in the previous analysis?
	1. Multivariable regressions
	2. Clustering analysis
	3. Other

## Data Source

The spatial data formulated in the previous step should be the starting place for this analysis. Additional data should be aggregated for explanatory variables identified in the first step of the analysis. These variables can be grouped by zip-code using a number of different approaches, including but not limited to:

* Median
* Mean
* Top/Bottom Quartile
* % of occurrence (for categorical data)

Potentially fruitful areas of exploration include

* Size and type of business
* Percentage of cost covered by incentive
* Program
* program administrator
* Savings levels
* Minority owned status

## Gaps in Participation by Socio-Demographic Groups and Geographic Area

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research****Question** | **Indicator** | **Measurement** | **Data Sources** | **Data****Analysis** |
| Are there gaps in program participation by socio-demographic groups? | Participation | Number of businesses participating by size, type of business, minority-owned business, etc. | PAs (Cedars), CEC, Census Bureau | GIS SpatialAnalysis informed by statistical analysis |

**Spatial Granularity:** Zip-code

**Time Frame**: Annual

**Variable(s) to Explore:** *TBD*

## Gaps in Spending by Socio-Demographic Groups and Geographic Area

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research****Question** | **Indicator** | **Measurement** | **Data Sources** | **Data****Analysis** |
| Are there gaps in program participation by socio-demographic groups? | Participation | Number of businesses participating by size, type of business, minority-owned business, etc. | PAs (Cedars), CEC, Census Bureau | GIS SpatialAnalysis informed by statistical analysis |

**Spatial Granularity:** Zip-code

**Time Frame**: Annual

**Variable(s) to Explore:** *TBD*

## Gaps in Savings by Socio-Demographic Groups and Geographic Area

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research****Question** | **Indicator** | **Measurement** | **Data Sources** | **Data****Analysis** |
| Are there gaps in program participation by socio-demographic groups? |  Energy Savings | kWh/MW/therms saved per business | PAs (Cedars), CEC | GIS SpatialAnalysis informed by statistical analysis |

**Spatial Granularity:** Zip-code

**Time Frame**: Annual

**Variable(s) to Explore:** *TBD*