**Infiltration Data Retrieval, Processing, and Validation Outline**

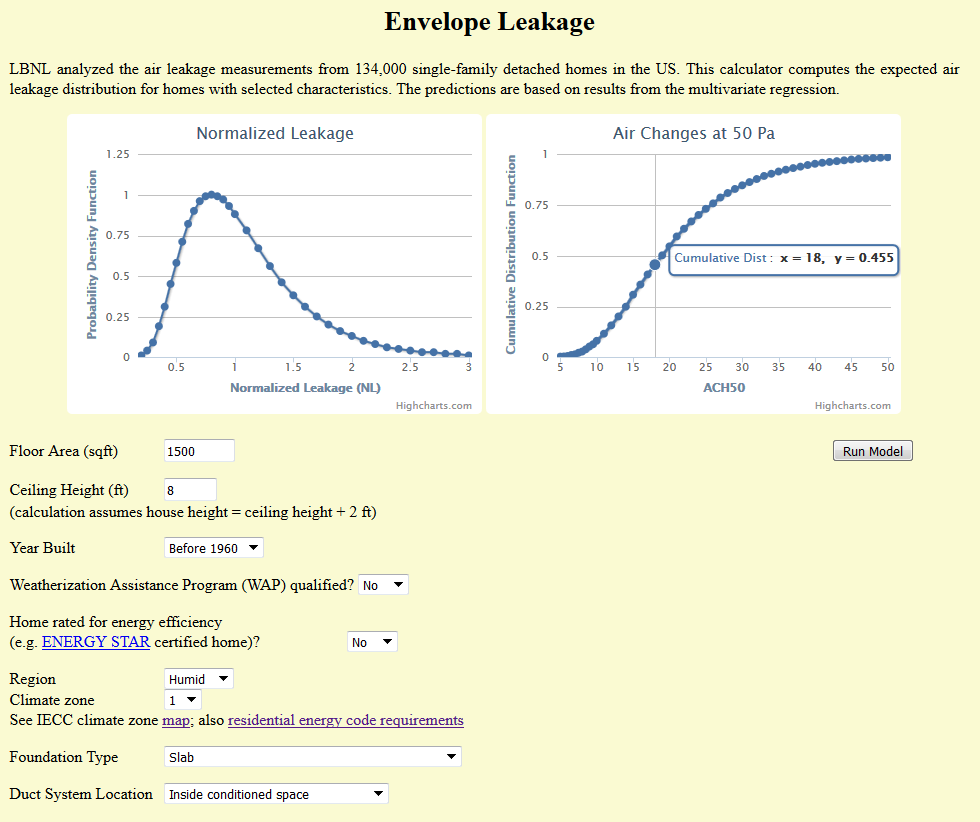
**John Alley**

**4/27/17**

# DATA GATHERING

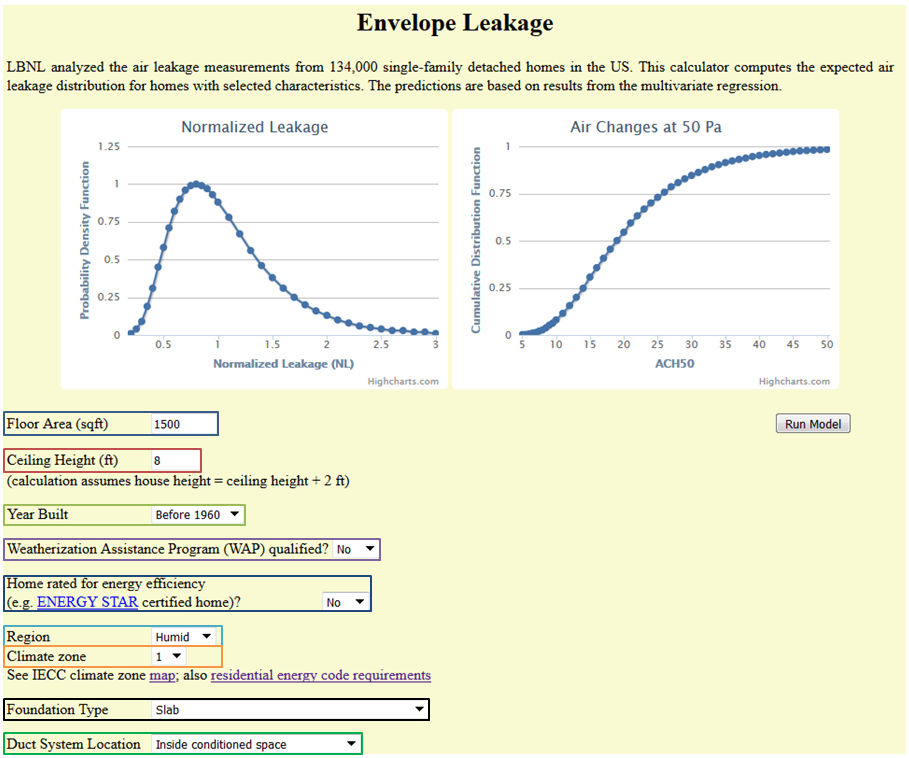
## The Laurence Berkeley Laboratory Model

1. Most comprehensive analyzation of air leakage measurements for single-family detached homes in the US ~134,000 single-family detached homes in the US
2. Created a model that takes 9 inputs and returns 2 Charts based on inputs:
   1. Normalized Leakage – Probability Density Function
   2. ACH50 – Cumulative Distribution Function



## Retrieving the Data

1. Compiled a list of ~32,000 combinations of homes to be evaluated using this tool
2. Used a script within Python to programmatically retrieve the data from the website
   1. Dictionary was created so that Text Inputs could be put into a URL Queries
      1. EXAMPLE: Colors in URL and Picture Match
         1. http://resdb.lbl.gov/main.php?step=2&sub=2&run\_env\_model=&dtype1=&dtype2=&is\_ca=&calc\_id=2&**floor\_area=1500**&**house\_height=8**&**year\_built=1**&**wap=0**&**ee\_home=0**&**region=1**&**zone=1**&**foundation=1**&duct=1



* 1. Sent requests to the website using the unique URL’s
  2. Read in graphs as list of (x,y) coordinates
  3. Paired coordinate lists with respective input values in a dataframe

1. Changed Numerical Inputs back to their text values: e.g. 6 🡪 2000’s or 1 🡪 Slab
2. Split Normalized Leakage and ACH50 into two separate dataframes
   1. Saved as .tsv files

## Manipulating the Data

### Cumulative Distribution to Probability Distribution

1. Separated x and y values for ACH50 into separate lists
2. Probability distribution created by making a new list of y-values using the following formula
   1. Let *n* be the *nth* element in [y0,y1,…,yn] and [Y0,Y1,…,Yn]. Where [y0,y1,…,yn] is a cumulative distribution of *y* and [Y0,Y1,…,Yn] is the probability distribution of *y*.   
      Yx+1  = yx+1 – yx for 0 < *x* < *n*
3. Set of columns created using x-values as headers
   1. Y-values populated across the columns

### Collapsing Variables

1. Total list of variables > 32,000 which is too much for a probability distribution.
   1. Needed to collapse variables with similar distributions
2. Summed up the mean and variance for each distribution using values from x and y distributions.
3. Looked at individual variables to see if they were distributed similarly under the Central Limit Theorem. 3 possible collapses were identified
   1. Ducts

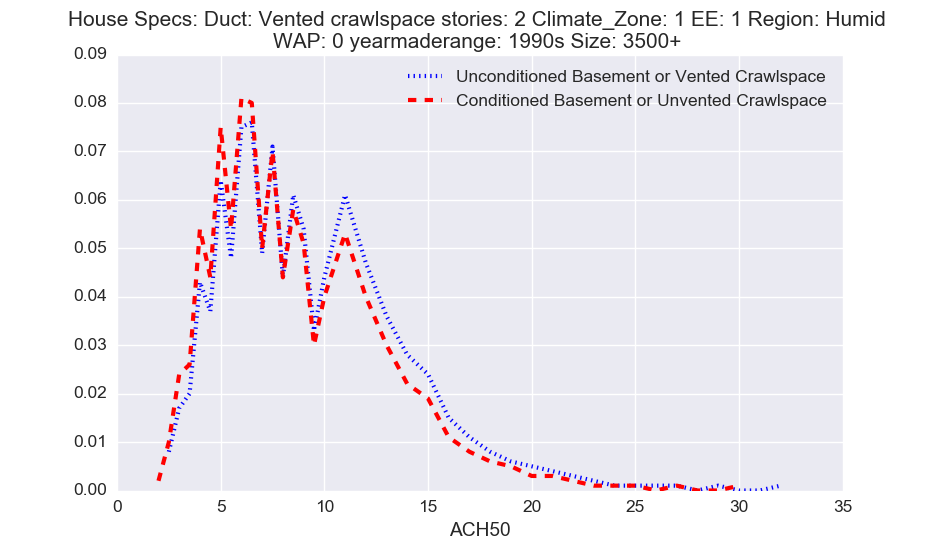
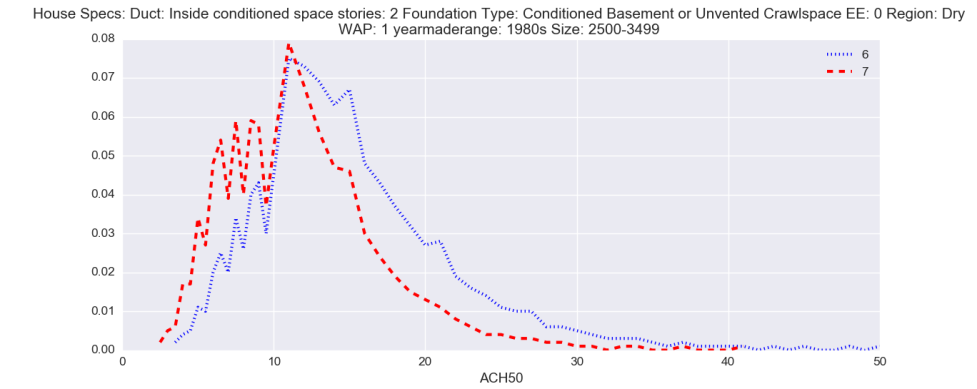
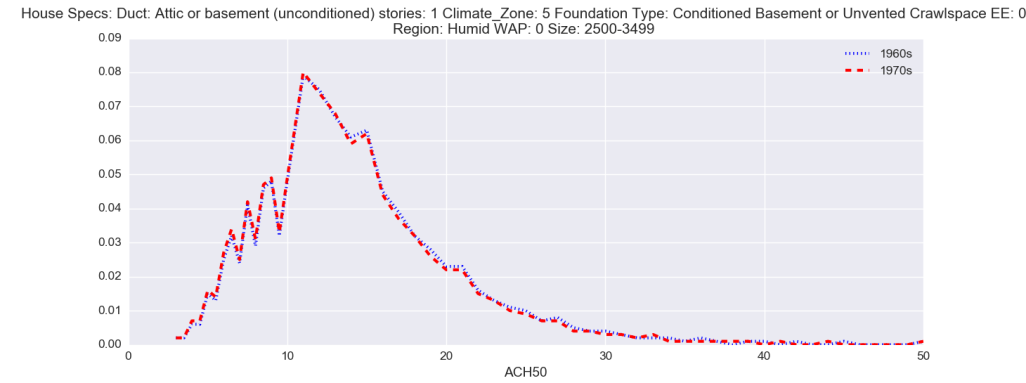
|  |  |  |  |
| --- | --- | --- | --- |
| **Foundation Type** | **Mean** | **stddev** | **Sum of Count** |
| Conditioned Basement or Unvented Crawlspace | 14.00438252 | 5.968563674 | 10368 |
| Unconditioned Basement or Vented Crawlspace | 14.96868875 | 6.092116677 | 10368 |
| **Grand Total** | **14.48653564** | 6.037066622 | **20736** |

* 1. Climate Zones

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Climate\_Zone** | **Region** | **Mean** | **Stddev** | **Count** |
| **6** |  | **12.45580208** | 5.571613 | **3456** |
|  | Dry | 12.56903096 | 5.596565 | 1728 |
|  | Humid | 12.34257321 | 5.546226 | 1728 |
| **7** |  | **12.49800145** | 5.581249 | **3456** |
|  | Alaska | 12.65342969 | 5.614761 | 1728 |
|  | Humid | 12.34257321 | 5.546226 | 1728 |

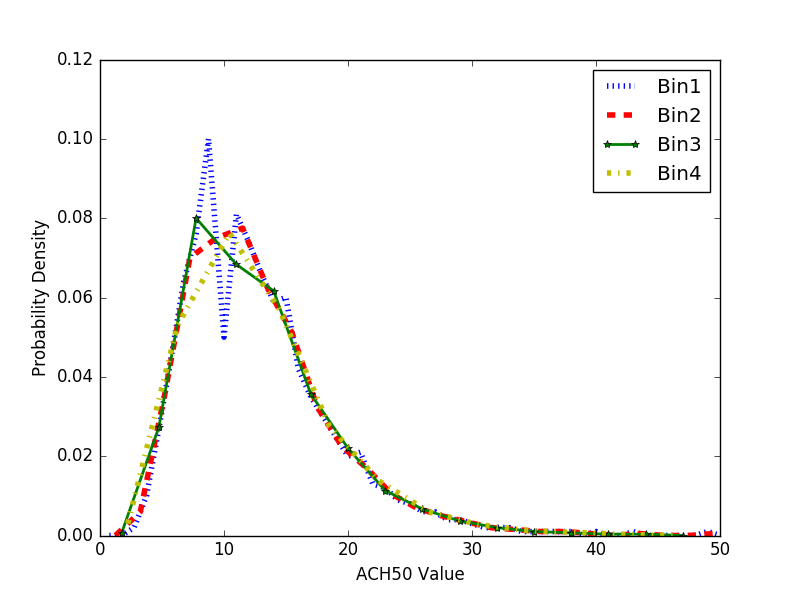
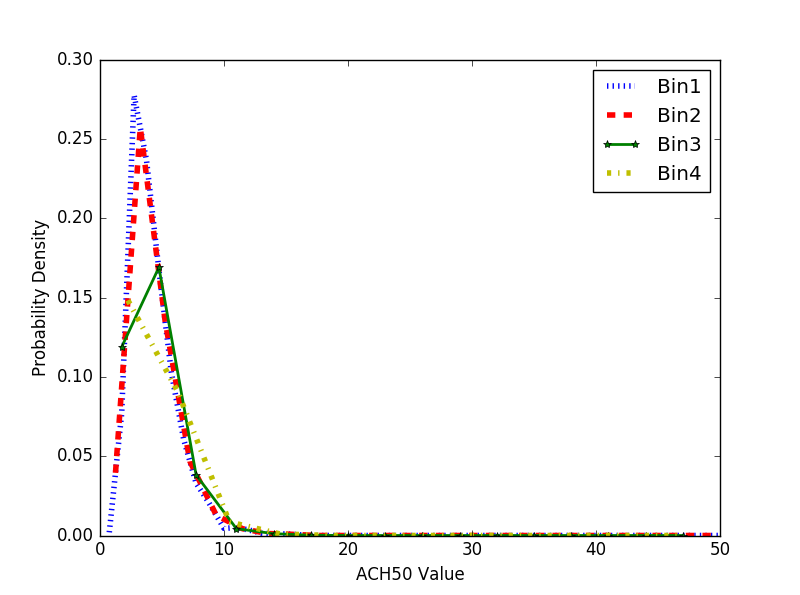
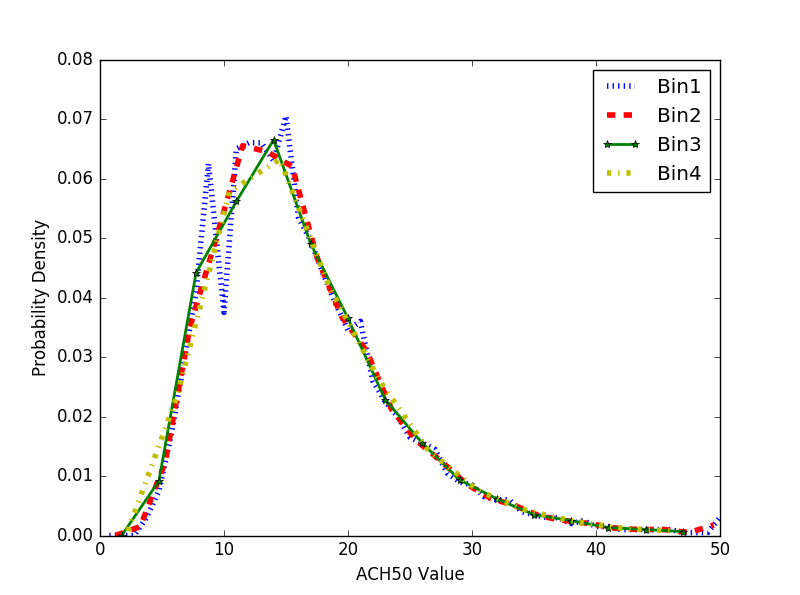
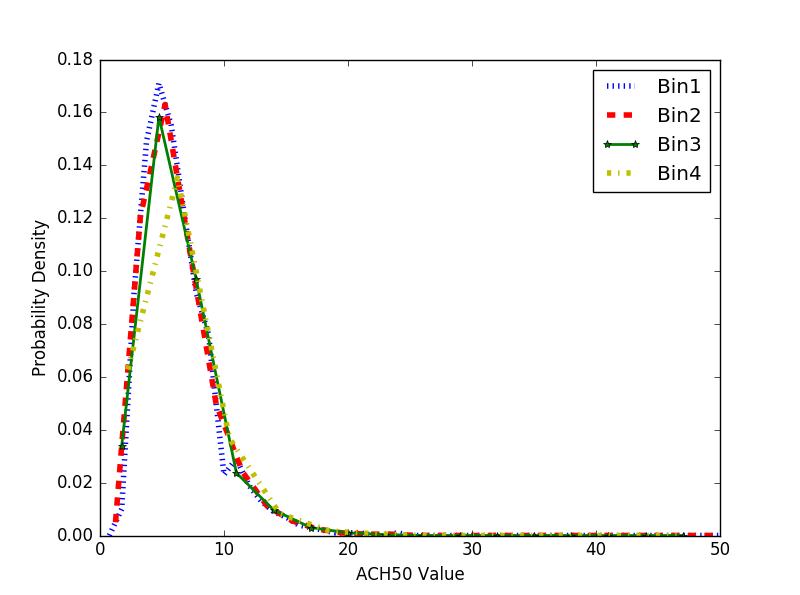
* 1. Vintage

|  |  |  |  |
| --- | --- | --- | --- |
| **yearmaderange** | **Mean** | **Standard Dev.** | **Count** |
| 1970s | 15.97570226 | 5.987082721 | 5184 |
| 1980s | 15.69571788 | 5.958146967 | 5184 |
| **Grand Total** | **15.83571007** | 5.973081055 | **10368** |

1. Since each of the distributions has means that differ by 0.04 ACH50 Value, and the number of values making up the distribution significant, we can assume by the central limit theorem that the distributions are similar and can be collapsed. Since our distributions depend highly on having granular data, we plotted randomly sampled houses with the only the variables above being different parameters.
   1. Foundation Type - ACCEPT  
      
   2. Climate Zones 6&7 - REJECT  
      
   3. Vintage 1960’s & 1970’ ACCEPT?  
      
2. By collapsing these 4 variables into 2 variables the number of rows was reduced by half.
   1. 'Foundation Type'
      1. 'Unconditioned Basement or Vented Crawlspace' & 'Conditioned Basement or Unvented Crawlspace' 🡪 ‘Basement or Crawlspace’
   2. ‘Vintage’
      1. ‘1960s’ & ‘1970s’ 🡪 ‘1960/70s’

### Binning the Data

1. Binned Data using the LBNL\_JSON\_pull.py file. Binned by 1,2,3,4 for values of ACH50.
2. Examined plots of binned data
3. Binning by 2 was best for smoothing irregularities in the data and still capturing the shape of the distribution



|  |
| --- |
|  |