WICC
Upstate Water Contamination Project

ADK OPINIONS ON ROAD SALT CONTAMINATION

EXECUTIVE SUMMARY

Water Insecurity Correction Coalition collected 125 surveys from residents across Adirondack Park to better understand resident attitudes toward road salt contamination and their water's safety. Survey results produced statistically significant conclusions. Most ADK residents surveyed did think that road salt contamination was a problem in their town. Respondents with private wells were more confident in their water's safety and more concerned with road salt contamination than respondents connected to public water supplies. 56% of respondents had never tested their household's water supply.

BACKGROUND

The analyzed dataset is the collected survey responses from 125 residents that span four different regions of the Adirondacks. Surveying was completed over the course of about a year (October 2022 to August 2023) by Water Insecurity Correction Coalition (WICC) via student volunteers for our Upstate Water Contamination Project to glean a strong understanding of Adirondack resident attitudes and knowledge levels regarding road salt contamination of their water supplies. This analysis has a two-pronged motive for a pair of audiences; a) policymakers and WICC's nonprofit partners, where the patterns ascertained from the survey data can be utilized to inform and improve programming around road salt contamination, and b) Adirondack residents, where conclusions drawn from this analysis will be presented at a community town hall event at the end of January to help inform and educate them about road salt contamination in their homes.

For context, de-icing salt is used on New York roads to mediate ice during the winter, but properties down-valley from these routes experience elevated levels of sodium and chloride contamination in their water. Experts say that as much as 77% of all surface water in the Adirondacks could be contaminated by road runoff [1]. Inherently, this problem significantly affects rural residents in the Adirondacks. Access to water quality testing is limited, since most households in the Adirondacks utilize private wells rather than regularly-tested municipal water connections. Residents' understanding of the problem is suspected to be lacking for the same reason – insufficient education on road salt

[1] Kelting, Daniel, et al. "Regional Analysis of the Effect of Paved Roads on Sodium and Chloride in Lakes." Adirondack Watershed Institute, Adirondack Watershed Institute, 11 Jan. 2022, www.adkwatershed.org/all-publications/regional-analysis-effect-paved-roads-sodium-chloride.

contamination, and disconnect from a larger, regulated water source. UWCP's motivating principle in collecting resident survey responses is that there is significant value in a) knowing whether or not your water is contaminated, and subsequently b) knowing what to do about it and keep yourself safe. The first step in achieving this is to find out what residents know, and if they care about what they know.

The goal of this data analysis was thus to find patterns of resident attitudes around road salt contamination (with the "lead" variable being their provided response to the question, "*On a scale of 1-5, which of the following best describes your opinion regarding road salt contamination affecting your town's drinking water supply*?" where 5 indicated that road salt, in their opinion, was a "significant problem," and 1 indicated their belief that it is "not a problem").

From a policy and programming standpoint, finding patterns of a) where in the Adirondacks do people care and/or know about road salt contamination, and b) who in the Adirondacks cares and/or knows about road salt contamination. The "where" spans from the four different survey tracts we collected responses from throughout the year, to more specifically the ZIP codes of where respondents lived. The "who" is a more demographically-framed question, wondering if people have differing opinions based on their water source (public or private water), their income levels, whether or not the lived on a street where salt is appled in the winter, or if whether or not they are confident in the safety of their household water suppy.

ASSUMPTIONS

Since WICC members both assisted in the collection of survey samples and completed the subsequent analysis of the data, we did have some assumptions about the general patterns of the responses – analysts went in believing that, on average, those with private wells would a) be less worried about road salt as a problem, b) be more confident in their water's safety, and c) have tested their water less frequently than their neighbors connected to the heavier-regulated public water supply. Though they were more confident in their water's safety, analysis surprisingly showed that those with private wells were more worried about road salt as a problem than their municipal-connected counterparts.

WICC members also completed some preliminary statistical analyses of this dataset (attached as Appendix A and B), having most notably found a statistically significant association between a respondent's road salt opinion and 1) confidence in their water's safety, 2) if road salt was applied to their street last winter, and 3) their ZIP code.

FINDINGS & NEXT STEPS

To make the most of this analysis, it is suggested that policy and programming recommendations be curated based on the data collected. These takeaways do have implications for populations that nonprofits should target educational and advocacy-oriented programming regarding road salt contamination. They also provide insight into what sub-populations of Adirondack residents warrant more resource provision (i.e. water quality testing).

1. A majority of Adirondack residents that responded to our survey do think that road salt contamination is a problem in their town.
2. Opinions on road salt vary significantly across different ZIP codes.
3. Respondents with private wells were more confident in their water's safety than respondents connected to public water supplies.
4. Respondents with private wells were slightly more concerned about road salt being a problem in their town than respondents connected to public water supplies.
5. Respondents who lived on a street where salt was applied last winter thought that road salt was a more significant problem in their town than respondents who had not had their road salted.
6. How recently a respondent's last water quality test was did not have a signifigant influence on their road salt opinion, nor on their confidence regarding their tap water's safety.

WICC
Upstate Water Contamination Project

Appendix A: AI-Assisted Statistical Analysis of ADK Survey Responses

ADK Data Analysis

Water Insecurity Correction Coalition

Produced by: Nicolas Gentile, Arianna Trapp, Annabel Gregg

Statistical Analysis of Confidence Level Associations

- Confidence Level and Income Category
    - P-Value: 0.405
    - Interpretation: No significant association between confidence levels and income categories.
- Confidence Level and Water Source
    - P-Value: < 0.001
    - Interpretation: Significant association found. Respondents' confidence levels are influenced by their water source.
- Confidence Level and Last Water Test
    - P-Value: 0.591
    - Interpretation: No significant association between confidence levels and the time since the last water test was conducted.
- Confidence Level and Road Salt Opinion
    - P-Value: < 0.001
    - Interpretation: Highly significant association found. The opinion on road salt strongly influences respondents' confidence levels.
- Confidence Level and Survey Tract
    - P-Value: 0.062
    - Interpretation: Marginal association detected. Suggests potential significance at a higher alpha level, indicating that survey tract may have an effect on confidence levels.
- Confidence Level and Whether Road Salt Was Applied
    - P-Value: 0.043
    - Interpretation: Significant association found. The application of road salt is associated with variations in confidence levels among respondents.

WICC
Upstate Water Contamination Project

Summary: The analysis reveals that environmental factors, such as water source and opinions on road salt, significantly affect confidence levels. In contrast, economic factors like income and practical measures, such as the timing of the last water test, appear not to be significant influencers. The marginal result for the survey tract suggests a geographical pattern that could be explored further.

```
from scipy.stats import chi2_contingency, f_oneway

# Function to perform Chi-square test
def chi_square_test(data, var1, var2):
    contingency_table = pd.crosstab(data[var1], data[var2])
    chi2, p, dof, expected = chi2_contingency(contingency_table)
    return p

# Function to perform ANOVA test
def anova_test(data, group_var, continuous_var):
    groups = data[group_var].unique()
    args = [data[data[group_var] == val][continuous_var] for val in groups]
    f_stat, p = f_oneway(*args)
    return p

# List to store results
results = []

# Confidence Level + Income Category
p_value_income = chi_square_test(data_df, 'Confidence Level', 'Income Category')
results.append(('Confidence Level + Income Category', p_value_income))

# Confidence Level + Water Source
p_value_water_source = chi_square_test(data_df, 'Confidence Level', 'Water Source')
results.append(('Confidence Level + Water Source', p_value_water_source))

# Confidence Level + Last Water Test
# This variable seems to be ordinal (having a natural order), so ANOVA could be used if
treated as continuous.
```

```
# However, given that the variable is coded with numbers that do not have equal intervals and include a '99' code for missing data,
# we should treat it as categorical and use Chi-square test.
p_value_last_water_test = chi_square_test(data_df, 'Confidence Level', 'Last Water Test\xa0')
results.append(('Confidence Level + Last Water Test', p_value_last_water_test))

# Confidence Level + Road Salt Opinion
p_value_salt_opinion = chi_square_test(data_df, 'Confidence Level', 'Road Salt Opinion')
results.append(('Confidence Level + Road Salt Opinion', p_value_salt_opinion))

# Confidence Level + Survey Tract
p_value_survey_tract = chi_square_test(data_df, 'Confidence Level', 'Survey Tract\xa0')
results.append(('Confidence Level + Survey Tract', p_value_survey_tract))

# Confidence Level + Whether Road Salt Was Applied
p_value_salt_applied = chi_square_test(data_df, 'Confidence Level', 'Salt applied to street last year\xa0')
results.append(('Confidence Level + Whether Road Salt Was Applied', p_value_salt_applied))

# Convert results to DataFrame for better readability
results_df = pd.DataFrame(results, columns=['Test', 'P-Value'])

results_df
```

Statistical Analysis of ZIP Code Associations

- ZIP Code and Income Category
  - P-Value: 0.447
  - Interpretation: No significant association between ZIP codes and income categories.
- ZIP Code and Water Source
  - P-Value: 0.559
  - Interpretation: No significant association between ZIP codes and water sources.
- ZIP Code and Last Water Test
  - P-Value: 0.801

    - Interpretation: No significant association between ZIP codes and the time since the last water test.
- ZIP Code and Whether Road Salt Was Applied
    - P-Value: 0.653
    - Interpretation: No significant association between ZIP codes and whether road salt was applied to their road.
- ZIP Code and Road Salt Opinion
    - P-Value: 0.017
    - Interpretation: Significant association found. Opinions on road salt vary significantly across different ZIP codes.

Summary: The ZIP code appears to have a significant relationship only with the opinion on road salt, indicating that regional differences may influence residents' opinions on this environmental issue. Other factors such as income, water source, time since the last water test, and road salt application do not show significant variation by ZIP code.

```
# Performing chi-square tests for independence between ZIP code (Survey Tract) and other variables.

# List to store new results
zip_code_results = []

# ZIP code + Income Category
p_value_zip_income = chi_square_test(data_df, 'Survey Tract\xa0', 'Income Category')
zip_code_results.append(('ZIP code + Income Category', p_value_zip_income))

# ZIP code + Water Source
p_value_zip_water_source = chi_square_test(data_df, 'Survey Tract\xa0', 'Water Source')
zip_code_results.append(('ZIP code + Water Source', p_value_zip_water_source))

# ZIP code + Last Water Test
p_value_zip_last_water_test = chi_square_test(data_df, 'Survey Tract\xa0', 'Last Water Test\xa0')
zip_code_results.append(('ZIP code + Last Water Test', p_value_zip_last_water_test))

# ZIP code + Whether Road Salt Was Applied
```

```
p_value_zip_salt_applied = chi_square_test(data_df, 'Survey Tract\xa0', 'Salt applied to street last year\xa0')
zip_code_results.append(('ZIP code + Whether Road Salt Was Applied', p_value_zip_salt_applied))

# ZIP code + Road Salt Opinion
p_value_zip_salt_opinion = chi_square_test(data_df, 'Survey Tract\xa0', 'Road Salt Opinion')
zip_code_results.append(('ZIP code + Road Salt Opinion', p_value_zip_salt_opinion))

# Convert new results to DataFrame for better readability
zip_code_results_df = pd.DataFrame(zip_code_results, columns=['Test', 'P-Value'])

zip_code_results_df
```

Statistical Analysis of Road Salt Opinion Associations

- Road Salt Opinion and Last Water Test
    - P-Value: 0.578
    - Interpretation: No significant association between road salt opinions and the time since the last water test.
- Road Salt Opinion and Water Source
    - P-Value: 0.216
    - Interpretation: No significant association between road salt opinions and water sources.
- Road Salt Opinion and Whether Road Salt Was Applied
    - P-Value: 0.031
    - Interpretation: Significant association found. Opinions on road salt are related to whether road salt was applied to their road.
- Road Salt Opinion and ZIP Code
    - P-Value: 0.017
    - Interpretation: Significant association found. Different ZIP codes have varying opinions on road salt, indicating regional differences in perspectives.
- Road Salt Opinion and Income
    - P-Value: 0.836
    - Interpretation: No significant association between road salt opinions and income categories.

- Road Salt Opinion and Confidence Level
  - P-Value: < 0.001
  - Interpretation: Highly significant association found. Road salt opinions are strongly linked to confidence levels in the dataset.

Summary: The analysis highlights a significant relationship between road salt opinions and both the application of road salt and the regional location (ZIP code). Additionally, there is a very strong association with confidence levels. Income and the time since the last water test do not appear to significantly affect road salt opinions.

```
# Performing chi-square tests for independence between Road Salt Opinion and other
variables.

# List to store new results
road_salt_opinion_results = []

# Road Salt Opinion + Last Water Test
p_value_salt_opinion_last_test = chi_square_test(data_df, 'Road Salt Opinion', 'Last Water
Test\xa0')
road_salt_opinion_results.append(('Road Salt Opinion + Last Water Test',
p_value_salt_opinion_last_test))

# Road Salt Opinion + Water Source
p_value_salt_opinion_water_source = chi_square_test(data_df, 'Road Salt Opinion', 'Water
Source')
road_salt_opinion_results.append(('Road Salt Opinion + Water Source',
p_value_salt_opinion_water_source))

# Road Salt Opinion + Whether Road Salt Was Applied
p_value_salt_opinion_salt_applied = chi_square_test(data_df, 'Road Salt Opinion', 'Salt
applied to street last year\xa0')
road_salt_opinion_results.append(('Road Salt Opinion + Whether Road Salt Was Applied',
p_value_salt_opinion_salt_applied))

# Road Salt Opinion + ZIP code
p_value_salt_opinion_zip = chi_square_test(data_df, 'Road Salt Opinion', 'Survey
Tract\xa0')
```

```
road_salt_opinion_results.append(('Road Salt Opinion + ZIP code', p_value_salt_opinion_zip))

# Road Salt Opinion + Income
p_value_salt_opinion_income = chi_square_test(data_df, 'Road Salt Opinion', 'Income Category')
road_salt_opinion_results.append(('Road Salt Opinion + Income', p_value_salt_opinion_income))

# Road Salt Opinion + Confidence Level
p_value_salt_opinion_confidence = chi_square_test(data_df, 'Road Salt Opinion', 'Confidence Level')
road_salt_opinion_results.append(('Road Salt Opinion + Confidence Level', p_value_salt_opinion_confidence))

# Convert new results to DataFrame for better readability
road_salt_opinion_results_df = pd.DataFrame(road_salt_opinion_results, columns=['Test', 'P-Value'])

road_salt_opinion_results_df
```

The ANOVA Test:

- To explore the relationship between Household Size and Income (using the midpoints for income categories) resulted in a p-value of approximately 0.0074.

Interpretation: This p-value indicates that there is a statistically significant difference in the average income midpoints among the different household sizes. In other words, household size appears to have a significant effect on the income level.

This result suggests that as the household size changes, there may be a corresponding change in the income level, which could be due to various economic factors such as the number of earners in the household or economies of scale. This finding could be of interest when considering household dynamics and economic status.

—

```
# To perform ANOVA for combining household size and income, we will first need to ensure that income is treated as a continuous variable.
# However, the 'Income Category' in our dataset is categorical. For the purpose of this ANOVA, we'll need to assign a representative value for each income category.
# For simplification, we could assign the midpoint of each income range to the categories. However, for the highest category and 'Prefer not to answer', we need a different approach.
# One common approach is to assign the highest category a value that is higher than the midpoint of the last specified range.
# For 'Prefer not to answer', we could either exclude these entries or impute them. For this analysis, we'll exclude them to avoid skewing the results.

# Mapping midpoints to income categories, excluding 'Prefer not to answer'
income_midpoints = {
   1: 12500,   # midpoint of $0-$24,999
   2: 50000,   # midpoint of $25,000-$74,999
   3: 112500,  # midpoint of $75,000-$149,999
   4: 175000   # a value higher than the midpoint of the last specified range
}

# Create a new column 'Income Midpoint' with the mapped values
data_df['Income Midpoint'] = data_df['Income Category'].map(income_midpoints)

# Exclude the 'Prefer not to answer' entries (coded as 99 in 'Income Category')
anova_df = data_df[data_df['Income Category'] != 99]

# Now we can perform ANOVA with 'Household Size' as the group variable and 'Income Midpoint' as the continuous variable.
anova_p_value = anova_test(anova_df, 'Household Size', 'Income Midpoint')

anova_p_value
—
```

Household Size and Water Source

- P-Value: 0.267
- Interpretation: There is no significant difference in household sizes across different water sources.

Household Size and Last Water Test

- P-Value: 0.576

- Interpretation: No significant difference in household sizes based on the time since the last water test.

Survey Tract and Income Category

- P-Value: 0.677
- Interpretation: No significant difference in income levels across different survey tracts.

Survey Tract and Household Size

- P-Value: 0.042
- Interpretation: There is a significant difference in household sizes across different survey tracts. This suggests that household size varies by geographic location.

These results indicate that only the geographic location (survey tract) has a significant impact on household size, while the other factors like water source, time since the last water test, and income category do not show significant variation in relation to household size. This information can be particularly useful for understanding demographic patterns and planning community resources or services.

—

```
# Re-importing necessary libraries and re-loading the dataset as the code execution state was reset
import pandas as pd
from scipy.stats import f_oneway

# Function to perform ANOVA test
def anova_test(data, group_var, continuous_var):
    groups = data[group_var].unique()
    args = [data[data[group_var] == val][continuous_var] for val in groups]
    f_stat, p = f_oneway(*args)
    return p

# Load the dataset
data_path = '/mnt/data/ASK data coded.csv'
data_df = pd.read_csv(data_path)

# Mapping midpoints to income categories, excluding 'Prefer not to answer'
```

```
income_midpoints = {
    1: 12500,   # midpoint of $0-$24,999
    2: 50000,   # midpoint of $25,000-$74,999
    3: 112500,  # midpoint of $75,000-$149,999
    4: 175000   # a value higher than the midpoint of the last specified range
}

# Create a new column 'Income Midpoint' with the mapped values
data_df['Income Midpoint'] = data_df['Income Category'].map(income_midpoints)

# Excluding 'Prefer not to answer' (99) from Income Category for this analysis
anova_df_income_tract = data_df[data_df['Income Category'] != 99]

# Performing ANOVA for the suggested combinations

# 1. Household Size and Water Source
anova_p_value_household_water = anova_test(data_df, 'Water Source', 'Household Size')

# 2. Household Size and Last Water Test
anova_p_value_household_last_test = anova_test(data_df, 'Last Water Test\xa0',
'Household Size')

# 3. Survey Tract and Income Category
anova_p_value_tract_income = anova_test(anova_df_income_tract, 'Survey Tract\xa0',
'Income Midpoint')

# 4. Survey Tract and Household Size
anova_p_value_tract_household = anova_test(data_df, 'Survey Tract\xa0', 'Household Size')

# Compile the results into a DataFrame
anova_results = pd.DataFrame([
    ('Household Size and Water Source', anova_p_value_household_water),
    ('Household Size and Last Water Test', anova_p_value_household_last_test),
    ('Survey Tract and Income Category', anova_p_value_tract_income),
    ('Survey Tract and Household Size', anova_p_value_tract_household)
], columns=['Test', 'P-Value'])
```

Anova_results

Regression Analysis:

- Dependent Variable (Y): 'Confidence Level' (since it's a key variable of interest and is numerical, making it suitable for linear regression).
- Independent Variables (X): 'Household Size', 'Income Category', 'Water Source', 'Road Salt Opinion', 'Survey Tract', 'Last Water Test', and 'Salt applied to street last year'.
- We will exclude 'Income Category' 99 from the regression since it represents 'Prefer not to answer'.
- For the 'Water Source' and 'Survey Tract' categorical variables, we'll create dummy variables for inclusion in the regression model.

OLS Regression Results

- R-squared: 0.989 - This suggests that 98.9% of the variability in the confidence level is explained by the model, which is an exceptionally high value and may suggest overfitting, especially if the data set is not very large.
- Adj. R-squared: 0.988 - This is adjusted for the number of predictors in the model and also indicates a very high proportion of explained variability.
- F-statistic: A very high F-statistic (1270) and its associated p-value close to zero indicates the model is statistically significant.
    - Coefficients:
    - Road Salt Opinion: The coefficient (0.9838) suggests a strong positive relationship with confidence level. For each one-unit increase in road salt opinion, there is a 0.9838 increase in the confidence level.
    - Water Source: The dummies for water source show varying impacts. 'WaterSource_3' and 'WaterSource_99' have negative and positive coefficients, respectively, suggesting different impacts on confidence level depending on the water source.
    - Survey Tract: Different survey tracts also show varying impacts on confidence level, with positive coefficients indicating higher confidence levels in certain tracts.

The p-values for individual coefficients indicate which variables are statistically significant. In this model:

- Road Salt Opinion has a highly significant effect on confidence level ($p < 0.001$).

- Water Source category 3 has a significant negative effect (p = 0.002), and category '99' (which may represent missing or other categories) has a significant positive effect (p = 0.016).
- Survey Tract shows significant differences in confidence levels across different tracts (p-values ranging from 0.037 to 0.000).
- Household Size does not have a significant effect on confidence level (p = 0.933), nor does the constant term of the model (p = 0.877).

Model Considerations:

- The high R-squared value may indicate that the model is too good to be true in real-world data, possibly due to overfitting.
- The significance of the '99' category for water source suggests that the handling of missing or unreported data can substantially affect the model, and thus, interpretations should be made with caution.

Overall, the model suggests that road salt opinion is a strong predictor of confidence level, with variations by water source and geographic location. This model can be further refined or validated with additional data to ensure robustness and to avoid overfitting.

```
import statsmodels.api as sm

# Preparing data for regression analysis
# Creating dummy variables for categorical predictors
water_source_dummies = pd.get_dummies(data_df['Water Source'], prefix='WaterSource', drop_first=True)
survey_tract_dummies = pd.get_dummies(data_df['Survey Tract\xa0'], prefix='SurveyTract', drop_first=True)

# Preparing the independent variables (X)
X = data_df[['Household Size', 'Road Salt Opinion']]
X = pd.concat([X, water_source_dummies, survey_tract_dummies], axis=1)

# Preparing the dependent variable (Y)
Y = data_df['Confidence Level']

# Adding a constant to the model (intercept)
X = sm.add_constant(X)
```

```
# Building the regression model
model = sm.OLS(Y, X).fit()

# Getting the summary of the regression model
model_summary = model.summary()
model_summary
```

Appendix B: Frequency Tables Generated Using STATA

last test of drinking water	Freq.	Percent	Cum.
within last 3 months	10	8.20	8.20
with last 3-6 months	4	3.28	11.48
within last 6-12 months	4	3.28	14.75
over 1 year ago	13	10.66	25.41
over 3 years ago	19	15.57	40.98
I have never tested my household drinki	72	59.02	100.00
Total	122	100.00	

tabulate salt_on_street

was road salt applied to your street last winter	Freq.	Percent	Cum.
No	16	13.11	13.11
Yes	80	65.57	78.69
Unsure/Maybe	26	21.31	100.00
Total	122	100.00	

WICC
Upstate Water Contamination Project

tabulate road_salt_op

opinion on road salt contamination	Freq.	Percent	Cum.
1	11	9.09	9.09
2	6	4.96	14.05
3	44	36.36	50.41
4	29	23.97	74.38
5	31	25.62	100.00
Total	121	100.00	

.

tabulate confidence

confidence in safety of tap water	Freq.	Percent	Cum.
1	6	4.96	4.96
2	14	11.57	16.53
2.5	2	1.65	18.18
3	19	15.70	33.88
4	46	38.02	71.90
5	34	28.10	100.00
Total	121	100.00	

tabulate resident

full time resident?	Freq.	Percent	Cum.
No	21	16.94	16.94
Yes	96	77.42	94.35
Other	7	5.65	100.00
Total	124	100.00	

tabulate source

drinking water source	Freq.	Percent	Cum.
I have a private well	51	41.13	41.13
I am connected to the towns public wate	62	50.00	91.13
I don't know	1	0.81	91.94
Other	10	8.06	100.00
Total	124	100.00	

Appendix C: Survey Distributed by WICC to ADK Residents