RESEARCH BRIEF

The Source You Trust

June 2022

Nedžad Arnautović

Heterogeneity of Office and Clerical Classifications

INTRODUCTION

Office and clerical classifications represent a significant portion of workers compensation (WC) payroll. Do risks that are entirely in the Office and Clerical industry group (IG) have similar loss experience to office and clerical employees working in businesses that belong to other IGs, such as Manufacturing or Contracting? Our research uncovers some interesting findings and primarily focuses on claim frequency experience for the largest, most reported exposure in WC: Class Code 8810—Clerical Office Employees Not Otherwise Classified (NOC).

KEY FINDINGS

- Claim frequency for policies having exposure for office and clerical classifications solely within the Office and Clerical IG is lower than claim frequency for policies having exposure for office and clerical classifications within other IGs
- For certain causes of injury, the difference in claim frequency is significantly lower
- Claim frequency for Class Code 8810 is nearly double when the governing class code is within the Goods and Services or Miscellaneous IGs as compared to policies for Class Code 8810 exposure only
- These claim frequency differences are persistent across states and policy sizes
- Cluster analysis identifies medical professionals, caregivers of the elderly, retail store employees, and building management as major contributors to the higher claim frequency differences
- Loss severity for Office and Clerical classifications is similar across all IGs, except for Contracting
- The results of the analysis are not sensitive to the choice of denominator for the frequency (payroll vs. premium) nor the types of claims analyzed (medical only vs. lost time)

STUDY DATA

Data used in this study is from NCCI's Unit Statistical Plan. NCCI collects, processes, and analyzes Unit Statistical Plan data for class ratemaking, experience rating production, actuarial analyses, and other NCCI thought leadership, products, and services.

For this study, we used Unit Statistical Plan experience for:

- Policy Years (PY) 2010–2019
- Thirty-seven jurisdictions where NCCI provides ratemaking services: AK, AL, AR, AZ, CO, CT, DC, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MD, ME, MO, MS, MT, NC, NE, NH, NM, NV, OK, OR, RI, SC, SD, TN, UT, VA, VT, and WV

TERMINOLOGY

Standard Exception Classification: Certain classes of employees in WC insurance who are common to many types of businesses and separately rated unless included specifically in the wording of the governing classification. These exceptions include clerical employees, drivers, and salespersons. These common occupations are not included in a basic classification unless specified in the classification wording.

Governing Classification: It is a basic classification, other than a standard exception classification, that produces the greatest amount of payroll on a policy. If a basic classification is not applicable, then the governing classification is the standard exception classification that is assigned the most payroll.

Class Code 8810: NCCI's ***Basic Manual*** states that Class Code 8810 duties must be limited to one or more of the following work activities:

- Creation or maintenance of:
    - Employer records
    - Correspondence
    - Computer programs
    - Files
- Telephone duties, including telephone sales
- Data entry or word processing
- Copy- or fax-machine operations, unless the insured is in the business of making copies or faxing for the public
- General office work similar in nature to the above

Multi-Industry Group (Multi-IG): Refers to the aggregated data for the office and clerical (O&C) classifications from policies that also have exposure in other IGs.

Office and Clerical Only (O&C Only): Refers to the aggregated data from O&C classifications from policies that have only O&C exposure (no exposure in any other IG).

Frequency: The number of claims observed at 1st report (i.e., 18 months) per $1 million of payroll. This facilitates comparisons because the frequency of O&C classifications is generally much lower than the frequency of non-O&C classifications.

Severity: Calculated as the sum of incurred indemnity plus medical amounts divided by the number of reported claims at 5th report.

Strain or Injury By: Cause-of-injury group aggregating several causes of injury into one. It includes injuries due to continual noise, twisting, jumping, holding or carrying, lifting, pushing or pulling, reaching, using tools or machinery, wielding or throwing, repetitive motion, and strains. For brevity purposes, the group is simply labeled as "Strain" throughout the article.

EXHIBITS

NCCI aggregates classifications into five broad IGs: Manufacturing, Contracting, Office and Clerical, Goods and Services, and Miscellaneous.[1] Figure 1 displays the distribution of classifications, payroll, premium, and claims by IG. We note that the Office and Clerical group accounts for 6% of total classifications, 64% of total payroll, 13% of total premium, and 16% of total claims. Given the large share of payroll it generates, it typically has the lowest claim frequency of any IG.

Figure 1: Distribution of Classifications, Payroll, Premium, and Claims

by Industry Group

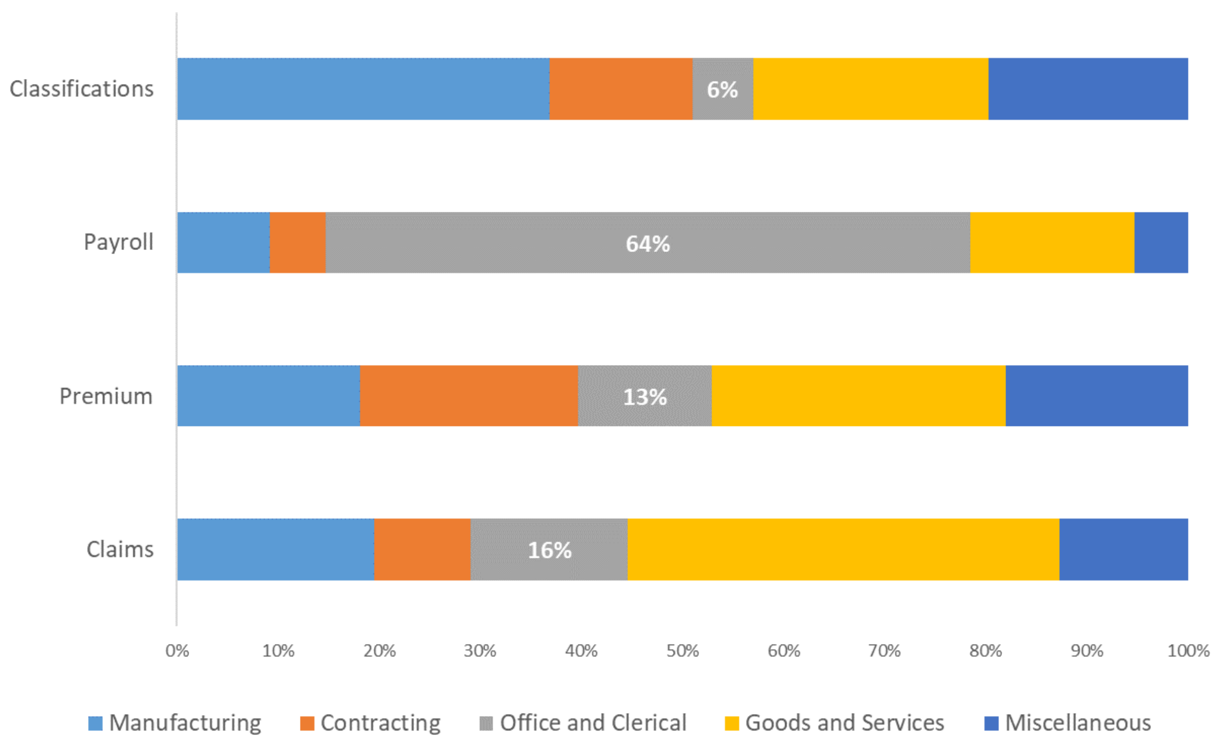


Figure 2 shows the three largest classifications in the Office and Clerical IG ranked by the amount of payroll observed during PY 2010–2019. Code 8810 comprises nearly 50% of Office and Clerical IG payroll and about 30% of the premium.

Figure 2: Top Three Classifications in the Office and Clerical Industry Group

Ranked by Amount of Payroll

Classification	Description	Payroll Share	Premium Share
8810	Clerical Office Employees NOC	48.6%	30.2%
8742	Salespersons or Collectors–Outside	11.8%	14.8%
8832	Physician and Clerical	8.2%	9.8%

[1] In Alaska, there is an additional IG—Oil and Gas. This IG was not included in the analysis.

Figures 3 through 5 display the relativity in frequency, severity, and share of claims between two O&C claim cohorts by cause of injury. Both cohorts represent aggregated O&C experience; however, the Multi-IG policies contain exposure from at least one non-O&C classification from other IGs. We define the relativity as the ratio of the Multi-IG value to its corresponding O&C Only value. The last column of each table shows whether the observed relativity is statistically significant at the 5% level. Here are several observations:

- All 10 cause-of-injury (COI) groupings have a frequency relativity greater than 1.15, and 5 out of 10 COI groupings are statistically significant at the 5% level.
- Likewise, 5 out of 10 COI groupings are statistically significant for claim severity.
- Unlike the frequency relativities, severity relativities are more similar between the two cohorts. Apart from "Caught in, Under, or Between," they range from 0.8 to 1.13.
- Both "Slip and Fall" and the "Strain" COI groupings are statistically significant in the frequency and severity domains.
- The "Slip and Fall" and "Strain" COI groupings account for approximately 50% of all O&C claims.

Figure 3: Frequency of Claims by Cause-of-Injury Groupings

O&C

Cause of Injury	Multi-IG	O&C Only	Relativity	Significant at 5%
Struck or Injured By	0.0298	0.0125	2.38	Yes
Rubbed or Abraded By	0.0006	0.0003	2.19	Yes
Burn	0.0036	0.0017	2.08	Yes
Strain	0.0440	0.0215	2.04	Yes
Slip and Fall	0.0519	0.0310	1.67	Yes
Miscellanous	0.0204	0.0113	1.81	
Caught in, Under, or Between	0.0054	0.0037	1.46	
Motor Vehicle	0.0056	0.0041	1.36	
Cut, Puncture, Scrape	0.0151	0.0120	1.26	
Striking Against or Stepping On	0.0091	0.0079	1.16	

Figure 4: Claim Severity by Cause-of-Injury Groupings

O&C

Cause of Injury	Multi-IG	O&C Only	Relativity	Significant at 5%
Caught in, Under, or Between	5,269	3,179	1.66	Yes
Slip and Fall	11,908	13,493	0.88	Yes
Strain	10,629	12,767	0.83	Yes
Motor Vehicle	25,015	31,310	0.80	Yes
Struck or Injured By	5,540	7,071	0.78	Yes
Cut, Puncture, Scrape	1,616	1,425	1.13	
Rubbed or Abraded By	9,732	8,989	1.08	
Burn	3,543	3,560	1.00	
Striking Against or Stepping On	4,516	5,209	0.87	
Miscellaneous	6,215	7,858	0.79	

Figure 5: Share of Claims by Cause-of-Injury Groupings

O&C

Cause of Injury	Multi-IG	O&C Only	Significant at 5%
Strain	23.8%	20.3%	Yes
Struck or Injured By	16.1%	11.8%	Yes
Striking Against or Stepping On	4.9%	7.4%	Yes
Caught in, Under, or Between	2.9%	3.5%	Yes
Burn	1.9%	1.6%	Yes
Slip and Fall	28.0%	29.3%	
Miscellaneous	11.0%	10.6%	
Cut, Puncture, Scrape	8.1%	11.3%	
Motor Vehicle	3.0%	3.9%	
Rubbed or Abraded By	0.3%	0.2%	

The data underlying the tables above include medical-only and indemnity claims. We obtained similar results when we restricted the analysis to lost-time claims. See the Appendix for more details.

Code 8810 contains almost 50% of payroll and 30% of the premium. To better understand the differences between the two cohorts, it is natural to examine this class code first.

Figure 6 displays claim frequency relativities of Code 8810 experience when a different governing classification exists on the policy versus the claim frequency of Code 8810 when the governing classification is Code 8810. When a governing classification is in either the Goods and Services or Miscellaneous IGs, the claim frequency of Code 8810 is about double the claim frequency than when the governing class code is Code 8810. Relativities for those two IGs are statistically significant at the 5% level.

Figure 6: Code 8810 Claim Frequency Relativities by IG of Policy Governing Class

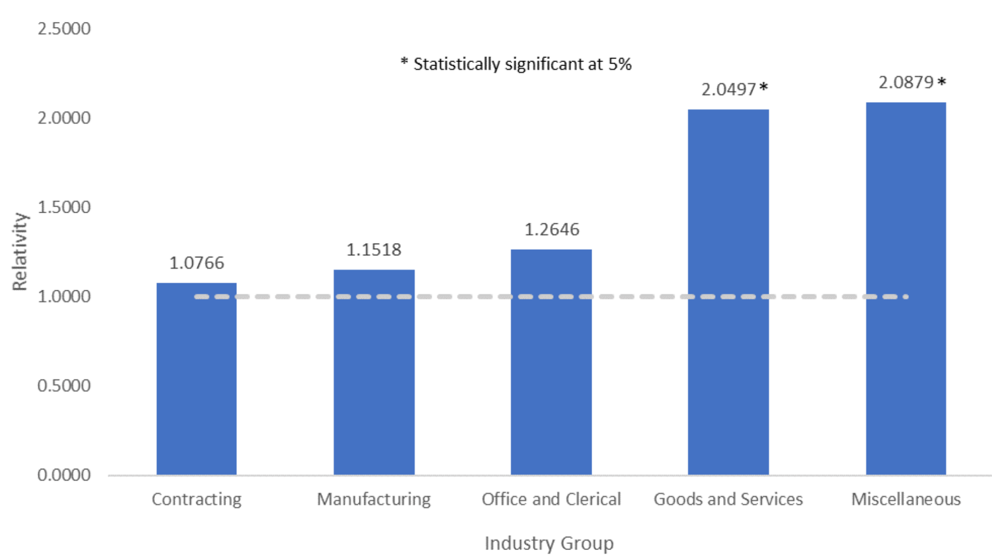
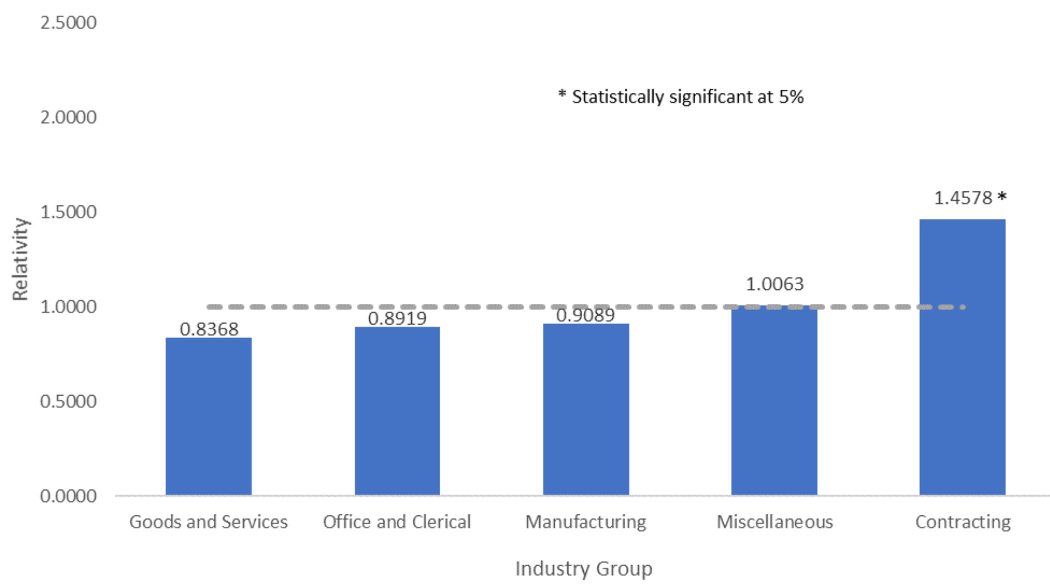


Figure 7 displays the severity relativities of Code 8810 experience when a different governing classification exists on the policy versus the claim frequency of Code 8810 when the governing classification is Code 8810. Except for Contracting, severities of Code 8810 from policies that have exposure in other IGs are rather similar.

Figure 7: Contracting Has the Highest Severity Relativity for Code 8810



In summary, workers assigned to Code 8810 are injured more often when the policy's governing class code is in the Goods and Services or Miscellaneous IGs. Although not shown in the article, these differences are evident for all states. In discussions with NCCI's Actuarial and Underwriting Committees, theories about Code 8810's observed frequency differences include:

- O&C workers from policies that have governing class codes in the Goods and Services or Miscellaneous IGs may temporarily perform non-O&C duties, which subjects them to greater risks on the job
- The lack of physical barriers between offices and service/production areas broadens the typical risk of injury for O&C workers and that may lead to more frequent injuries—especially slips and falls

Cluster Analysis

Cluster analysis is a technique in which a set of objects with similar characteristics are grouped together. We used this technique to identify governing class codes that contribute to frequency differences observed across IGs in Figure 6. We chose cluster analysis over other options because it relies on well-known algorithms and can produce goodness-of-fit statistics. This may not be the case with other methods.

We performed cluster analysis in five-dimensional space (R^5), meaning data points to be clustered are vectors with five components. For any governing class code, except Code 8810, each component is a loss frequency for Code 8810 experience representing one of the following causes of injury:

- Cut, Puncture, Scrape
- Slip and Fall
- Motor Vehicle
- Strain or Injury By
- Struck or Injured By

We selected the five groups because they account for 80% of all claims for Code 8810. We performed clustering in dimensions greater than five as well, and the results were similar.

Here is an example illustrating the data vector. For policies where the governing class code is Code 0005—Farm—Nursery Employees and Drivers, assume the following Code 8810 experience:

- 10 claims in Cut, Puncture, Scrape
- 30 claims in Slip and Fall
- 2 claims in Motor Vehicle
- 20 claims in Strain or Injury By
- 3 claims in Struck or Injured By
- $100 million dollars in payroll

Keeping in mind that we compute the frequency per $1 million of payroll, the five-dimensional loss-frequency vector used by the clustering algorithm would be $\langle \frac{10}{100}, \frac{30}{100}, \frac{2}{100}, \frac{20}{100}, \frac{3}{100} \rangle$. This construction is repeated for each governing class code to obtain 641 rows, as illustrated in Figure 8.

Figure 8: Illustration of Data Set Used by the Clustering Algorithm

All numbers are hypothetical

(1) Row	(2) Governing Class Code	(3) Description	(4) Governing Class Code Industry Group	(5) Cut	(6) Slip and Fall	(7) Motor Vehicle	(8) Strain	(9) Struck By	(10) Payroll
1	0005	Farm Nursery	Goods and Services	0.10	0.30	0.02	0.20	0.03	$100M
2	7720	Police Officers	Miscellaneous	0.08	0.25	0.05	0.18	0.07	$80M
…	…	…	…	…	…	…	…	…	…
641	3220	Can Manufacturing	Manufacturing	0.01	0.01	0	1	0	$5M

Before proceeding with the cluster analysis, we addressed the following issues:

- Governing class codes having a small volume of claims
- Correlations between the five cause-of-injury components

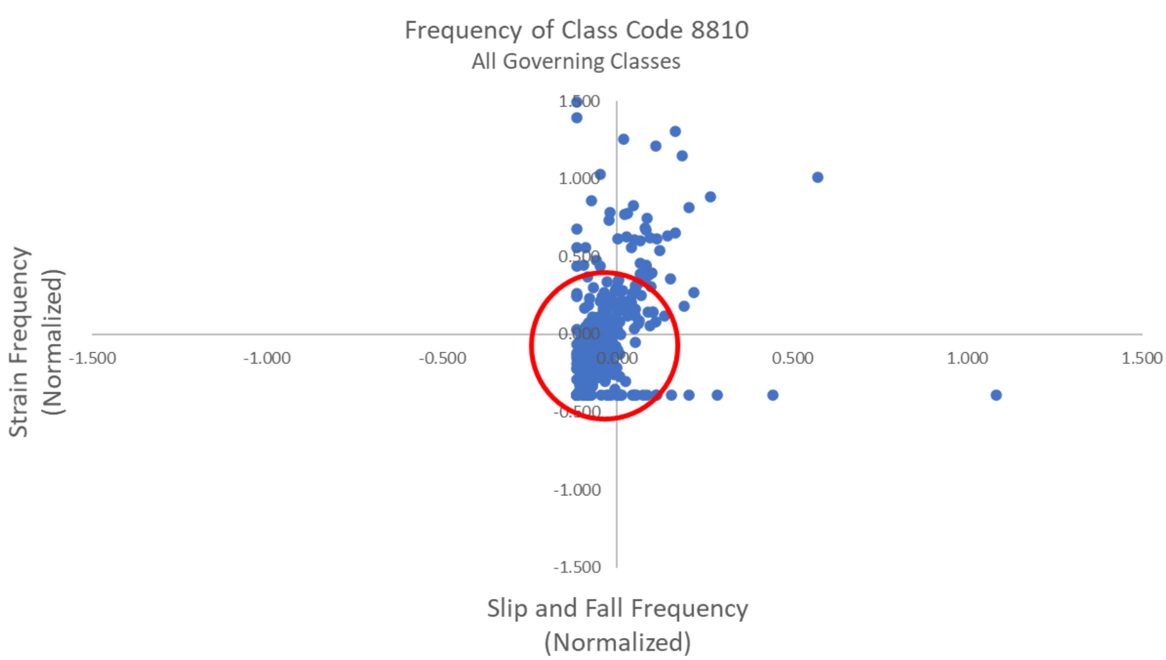
Handling of Governing Class Codes With Small Volumes of Claims

Figure 9 illustrates issues with including governing class codes that have a small volume of claims. Each point on the graph is an aggregation of Code 8810 experience across all policies and policy years with the same governing class code. Each point also represents the normalized[2] claim frequency plotted for two common causes of injury: slips and falls versus strains. Governing class codes with small volumes of O&C experience drive three observations:

1. Most governing class codes are grouped in one blob (indicated by the red circle), which makes the cluster analysis difficult to perform.
2. Regardless of which clustering algorithm was employed – K-Means, Gaussian Mixed Models, or Density Based Scans – we determined the optimal number of clusters to be one.
3. Some governing class codes have no claims or very few claims. The resulting edges along the axes of the graph also present challenges.

Figure 9: Strain vs. Slip and Fall Frequency

Code 8810 Experience for All Governing Class Codes



[2] A normalized random variable, Z, is one having a mean of zero and a standard deviation of one. That is, $Z = \frac{X-E[X]}{SD[X]}$ where E[X] and SD[X] are the expected value and standard deviation of non-normalized random variable X, respectively.

To address the issues, we identified the largest governing class codes representing approximately 90% of Code 8810's payroll. The remaining governing class codes were grouped into an "Other" category, one for each IG, and included in the cluster analysis. This approach reduced the number of five-dimensional vectors used in the cluster analysis from 641 to 163.

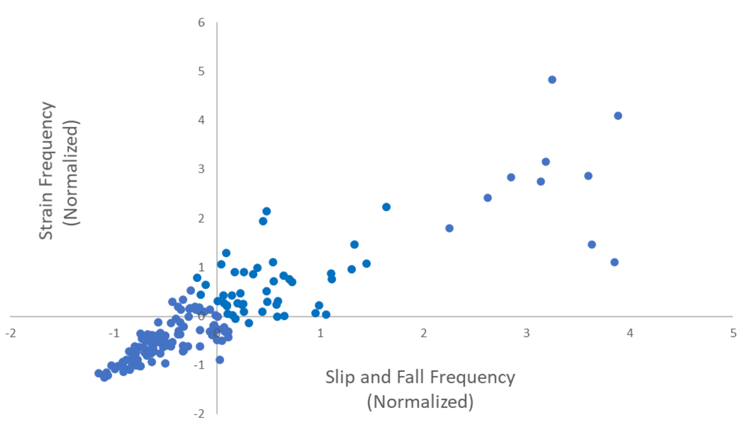
Correlations Between the Five Cause-of-Injury Components

Figure 10 illustrates the same analysis as shown in Figure 9, except for one major difference—we have reduced the number of governing classes. The one cluster blob described earlier is no longer present, but the two cause-of-injury components now show a high degree of correlation. The existence of such correlation may be puzzling at first but can be explained as follows:

- Each point in the graph is an aggregation of Code 8810 experience across all policies and policy years that have the same governing class code.
- Large insureds (more than $25 million in payroll) exhibit correlation between the two components, which may not be surprising since the O&C workers may perform non-O&C duties leading to such injuries.
- Insureds with smaller volume (below $25 million in payroll) exhibit negative correlation as they may often be claim-free in a given policy year. However, upon aggregation of the high-volume, small-risk experience, the correlation between the two causes of injury also becomes positive—similar to the large-risk experience.

Figure 10: Strain vs. Slip and Fall Frequency–Class Code 8810 Experience

Largest Governing Classes Representing 90% of Payroll



The existence of pairwise correlations between the five components of the clustering vector can lead to suboptimal results. Some goodness-of-fit statistics, which determine the most optimal number of clusters, are not applicable to data that has correlated components [3].

To remove the correlations between the components, we apply Principal Component Analysis (PCA) prior to clustering. Not only can PCA remove the correlations between the components and enhance the cluster analysis, but it can also reduce complicated data sets to a lower dimension and reveal hidden dynamics underlying the data set. For more details on PCA, reference the Appendix.

Figures 11 and 12 show three clusters based on the previous data set. Figure 11 does not apply PCA, while Figure 12 applies PCA prior to using the clustering algorithm. PCA transformed the original coordinate system in which the two components were correlated into a new coordinate system where there is no correlation between the two causes of injury. The new coordinate system also provides more dispersion between the data points. This allows the clustering algorithm to better identify clusters.

Figure 11: Strain vs. Slip and Fall Frequency—Class Code 8810 Experience
Largest Governing Classes Representing 90% of Payroll
Before PCA

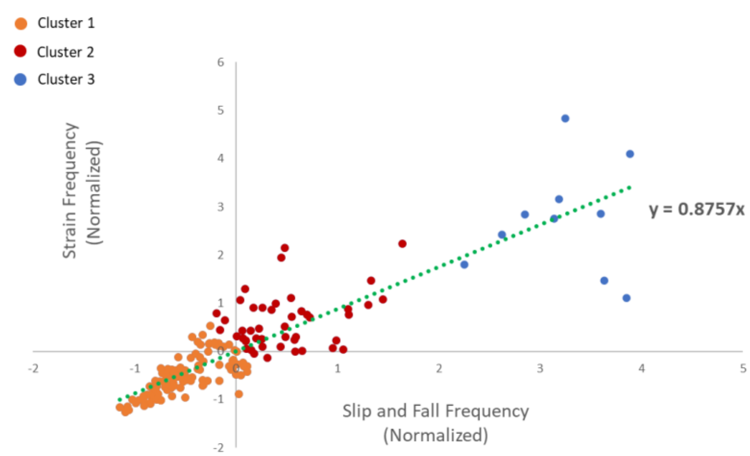
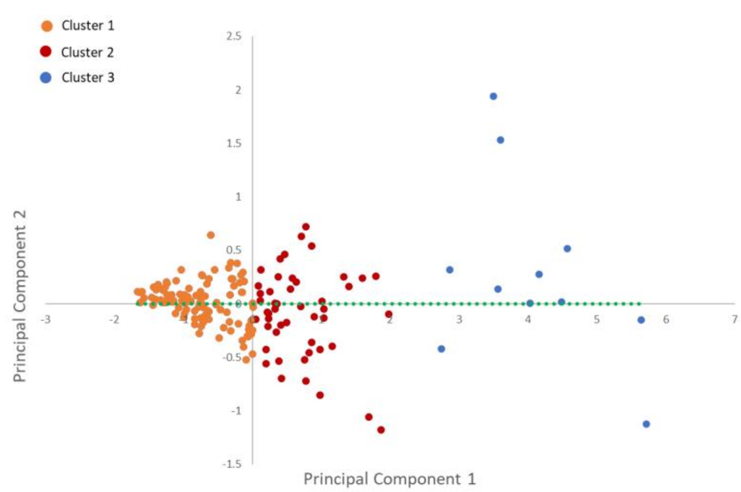


Figure 12: Strain vs. Slip and Fall Frequency—Class Code 8810 Experience
Largest Governing Classes Representing 90% of Payroll
After PCA



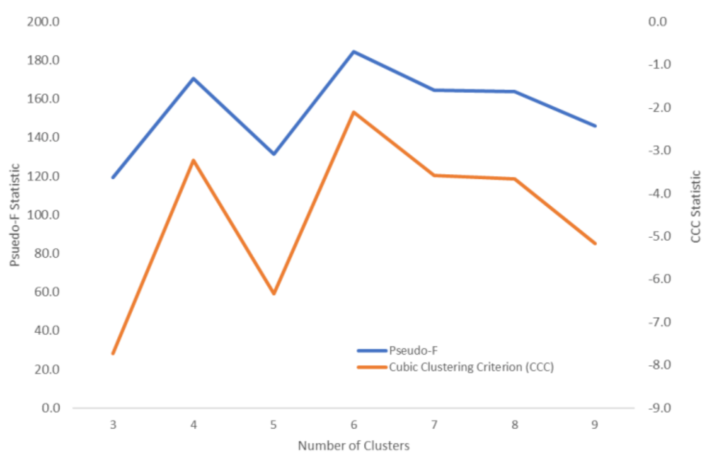
We selected weighted K-Means clustering with a Euclidean metric as our clustering algorithm, with payroll amounts for Code 8810 used as weights.

We used 10 policy years in the analysis. Restricting the analysis to fewer policy years and choosing different policy years in the formation of the governing class code data set did not yield different results in cluster assignments. See the Appendix for more details.

We used two test statistics to help determine the optimal number of clusters – Pseudo-F and Cubic Clustering Criterion. For a predetermined number of clusters, the procedure calculates these two statistics. We then repeated the process for a number of clusters between three and nine, and the number of clusters for which the two statistics are maximized was chosen as optimal. Figure 13 shows that both Pseudo-F and Cubic Clustering Criterion indicate the optimal number of clusters to be six. See [3] for further details about K-Means clustering as well as the two test statistics.

Figure 13: Optimal Number of Clusters

Pseudo-F and Cubic Clustering Criterion



The choice of clustering metric influences the optimal number of clusters shown in Figure 13. Euclidean metric, or L^2, is more sensitive to outliers than L^1. If there are outliers present in the data, both goodness-of-fit statistics will be suboptimal when the desired number of clusters is low and we select the Euclidean metric. For brevity purposes, we combined three clusters, each containing at most two outlying governing class codes and very small amounts of payroll, into Cluster 3—the cluster that exhibits the highest mean frequency. Hence, the final number of clusters shown is three and this manual "assignment" of governing class codes into three clusters differs slightly from the assignment selected by the K-Means unsupervised classification algorithm. Having Cluster 3 absorb few outlying governing class codes does not influence the conclusion of this analysis.

Figure 14 shows the cluster mean (i.e., average) for each of the three final clusters and for the two major causes of injury – Strain and Slip and Fall. The black dot on the chart shows the mean frequency for each of the two causes of injury for policies for which the governing class code is Code 8810. While the mean of Cluster 1 is similar to the mean when the governing class code is Code 8810, the same cannot be said for the means of Cluster 2 and Cluster 3. The mean of Cluster 2 is about twice the mean of governing Class Code 8810, and Cluster 3 is about four times higher than Cluster 1. We show the top five governing class codes comprising each cluster later in the article.

Figure 14: Strain vs. Slip and Fall—Mean Frequency for Clusters 1-3

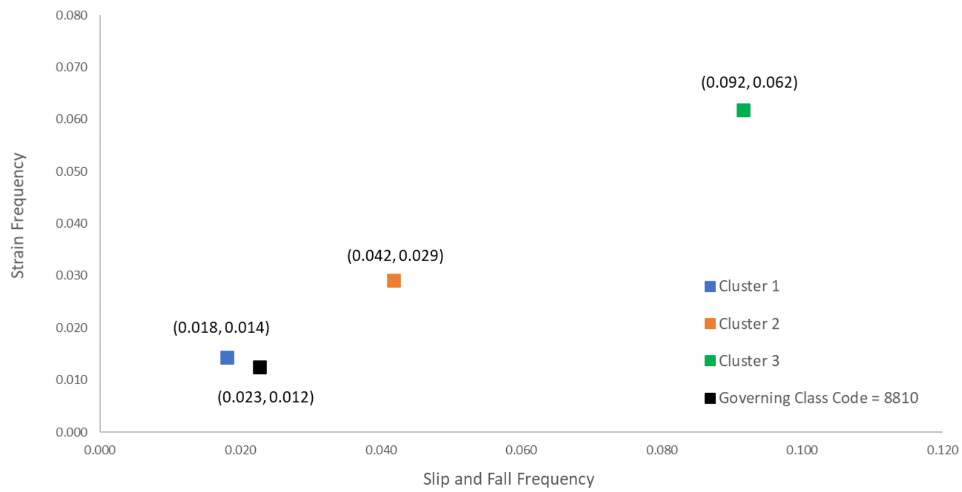


Figure 15 illustrates the distribution of payroll across the three clusters and IGs:

- Cluster 1 is a group of policies that have governing class codes for which Code 8810 loss frequency is lowest
- Cluster 2 is a group of policies that have governing class codes for which Code 8810 loss frequency is mid-range
- Cluster 3 is a group of policies that have governing class codes for which Code 8810 loss frequency is highest

As mentioned earlier, we did not include policies for which the governing class code is Code 8810 in the Office and Clerical IG codes.

Figure 15: Distribution of Code 8810 Payroll by Cluster and Industry Group

	Industry Group of the Governing Class Code				
Cluster	Manufacturing	Contracting	Office and Clerical	Goods and Services	Miscellaneous
1	24.7%	7.0%	13.8%	11.1%	2.6%
2	1.3%	0.6%	7.8%	22.8%	4.3%
3	0.0%	0.0%	0.0%	3.0%	1.0%

Figure 15 shows that the majority of Code 8810 payroll is concentrated in Cluster 1, when the governing class code is in Manufacturing, Contracting, or O&C. In Clusters 2 and 3, the majority of Code 8810 payroll is in the Goods and Services or Miscellaneous IGs, which have relatively higher claim frequencies as shown earlier.

Top Five Governing Class Codes by Cluster

Figures 16A and 17A display the top five governing class codes ranked by the amount of payroll in Code 8810 for each cluster. Figure 18A shows all governing class codes contained in Cluster 3, given it has the highest claim frequency classes. Governing class codes are depicted by oval shapes, while the mean of each cluster is depicted with a rectangular shape and in the same color as its governing class codes. For reference purposes, the means of the remaining two clusters are colored in gray. Below each figure is a table providing other details about the governing class codes, such as their IG and hazard group. We note the following:

- Most governing class codes in higher frequency clusters belong to the Goods and Services or Miscellaneous IGs
- The slip and fall cause of injury consistently has a higher claim frequency than strains
- Several governing class codes revolve around providing medical services to the public, such as:
    - Hospital—Professional Employees (Code 8833) in Cluster 2
    - Retirement Living Centers, Nursing Homes and Group Homes (Code 8842) in Cluster 3

Figure 16A: Strain vs. Slip and Fall Frequency—Cluster 1

Top Five Governing Class Codes

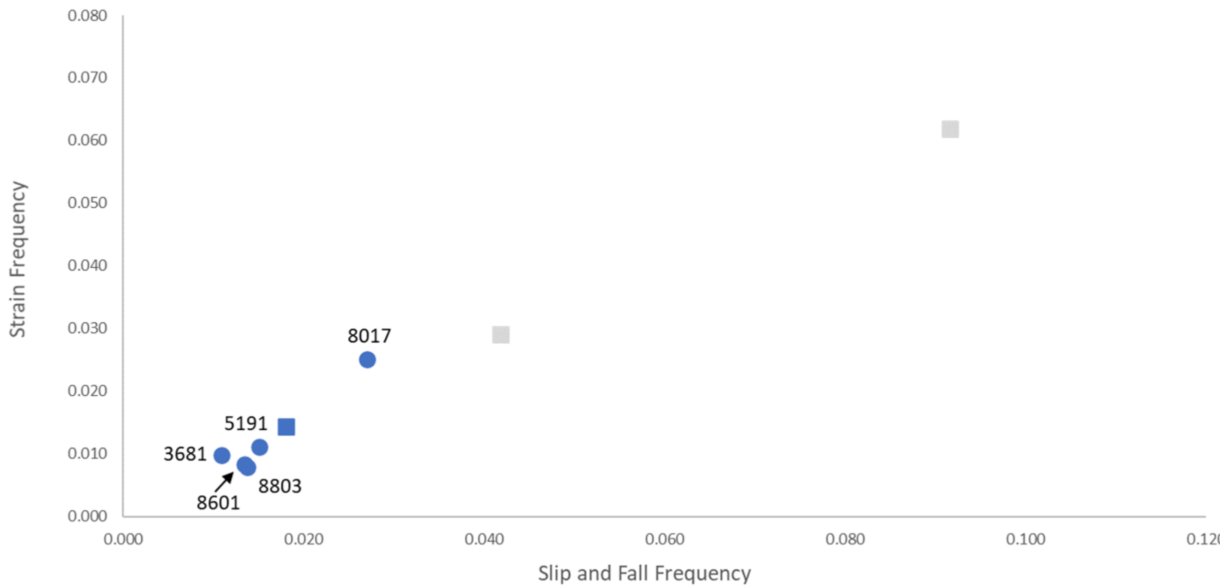


Figure 16B: Top Five Governing Class Codes in Cluster 1

Additional Details

Governing Class Code	Description	Industry Group	Hazard Group	Slip and Fall Frequency	Strain Frequency
3681	Television, Radio, Telephone or Telecommunication Device Manufacturing	Manufacturing	C	0.011	0.010
5191	Office Machine Installation, Inspection, Adjustment, or Repair	Goods And Services	E	0.015	0.011
8017	Store: Retail	Goods And Services	B	0.027	0.025
8601	Architectural or Engineering Firm—Including Salespersons and Drivers	Office and Clerical	F	0.013	0.008
8803	Auditor, Accountant, or Computer System Designer or Programmer—Traveling	Office and Clerical	E	0.014	0.008

Figure 17A: Strain vs. Slip and Fall Frequency—Cluster 2

Top Five Governing Class Codes

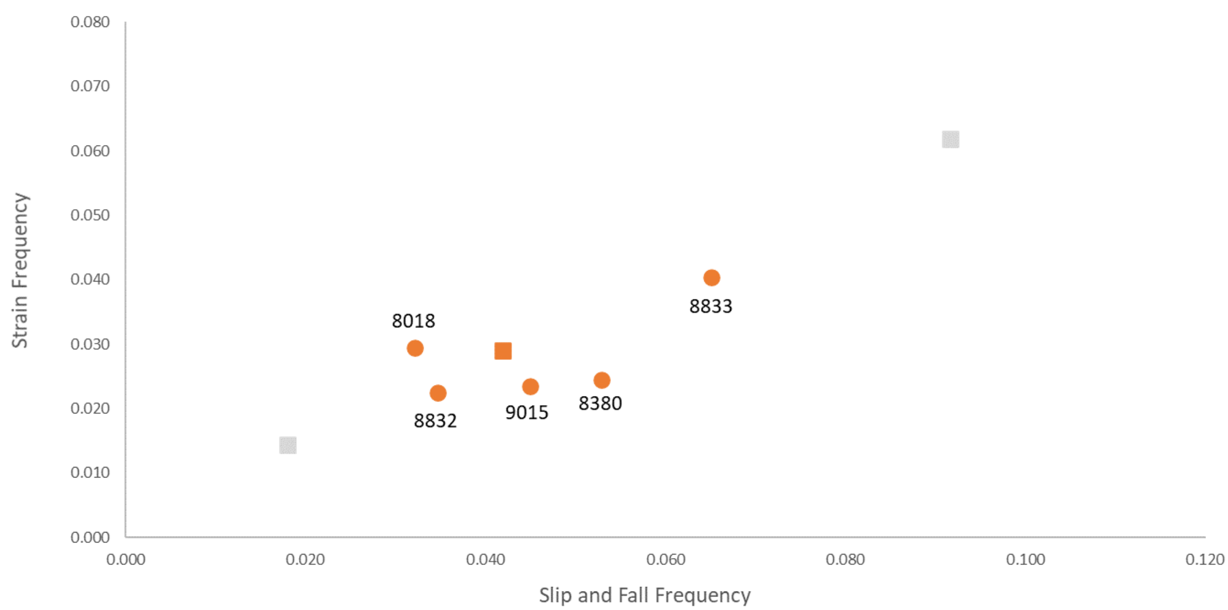


Figure 17B: Top Five Governing Class Codes in Cluster 2

Additional Details

Governing Class Code	Description	Industry Group	Hazard Group	Slip and Fall Frequency	Strain Frequency
8018	Store: Wholesale	Goods And Services	C	0.032	0.029
8380	Automobile Service or Repair Center and Drivers	Goods And Services	D	0.053	0.024
8832	Physician and Clerical	Office and Clerical	C	0.035	0.022
8833	Hospital—Professional Employees	Office and Clerical	C	0.065	0.040
9015	Building or Property Management—All Other Employees	Goods And Services	D	0.045	0.024

Figure 18A: Strain vs. Slip and Fall Frequency—Cluster 3
All Governing Class Codes

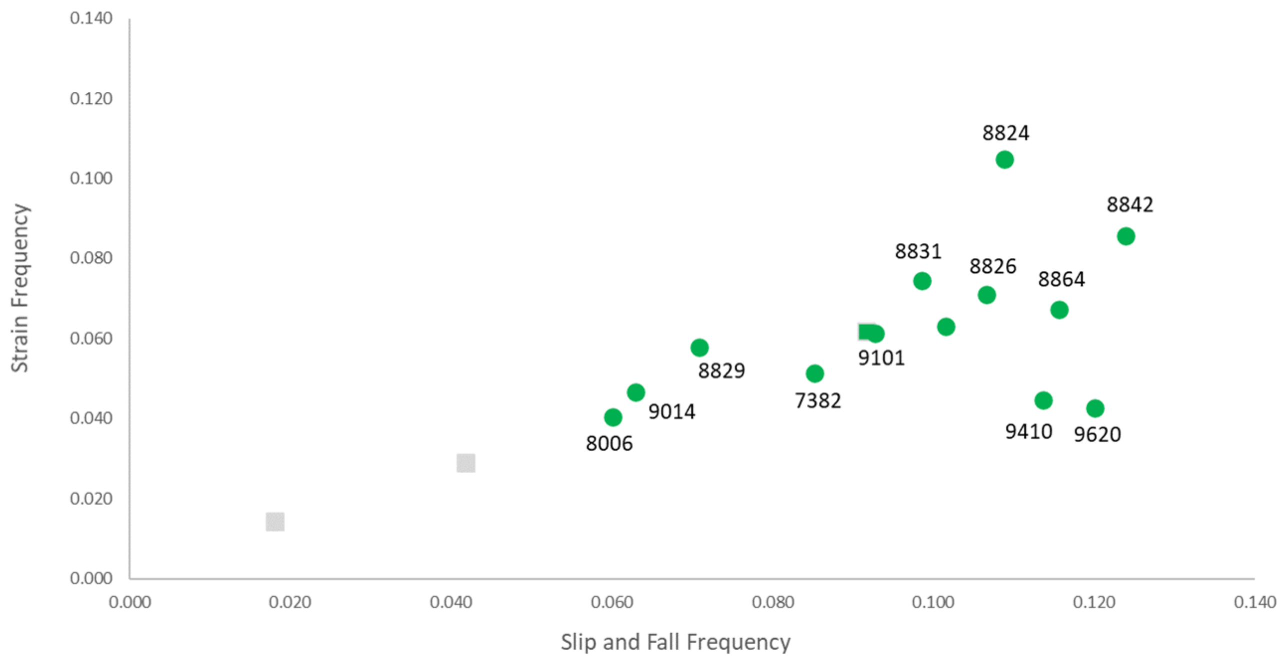


Figure 18B: All Governing Class Codes in Cluster 3
Additional Details

Governing Class Code	Description	Industry Group	Hazard Group	Slip and Fall Frequency	Strain Frequency
7382	Bus Co.—All Other Employees and Drivers	Miscellaneouos	D	0.085	0.051
7720	Police Officers and Drivers	Miscellaneouos	E	0.101	0.063
8006	Store—Grocery—Retail	Goods And Services	B	0.060	0.040
8824	Retirement Living Centers—Health Care Employees	Goods And Services	A	0.109	0.105
8826	Retirement Living Centers—All Other Employees and Salespersons, Drivers	Goods And Services	B	0.107	0.071
8829	Convalescent or Nursing Home—All Employees	Goods And Services	B	0.071	0.058
8831	Hospital—Veterinary and Drivers	Goods And Services	A	0.099	0.074
8842	Group Homes—All Employees and Salespersons, Drivers	Goods And Services	A	0.124	0.086
8864	Social Service Organization—All Employees and Salespersons, Drivers	Goods And Services	B	0.116	0.067
9014	Janitorial Services By Contractors—No Window Cleaning Above Ground Level and Drivers	Goods And Services	C	0.063	0.047
9101	College—All Other Employees	Goods And Services	B	0.093	0.061
9410	Municipal, Township, County, or State Employee	Goods And Services	C	0.114	0.045
9620	Funeral Director and Drivers	Goods And Services	E	0.120	0.043

Figure 19 shows strain versus slip and fall claim frequency for the aggregation of the smallest governing class codes into “Other” buckets for each IG. The Goods and Services and Miscellaneous “Other” buckets have higher frequencies than Manufacturing or Contracting. No “Other” bucket had a claim frequency high enough to belong to Cluster 3.

Figure 19: Strain Frequency vs. Slip and Fall Frequency for IG “Other” Buckets

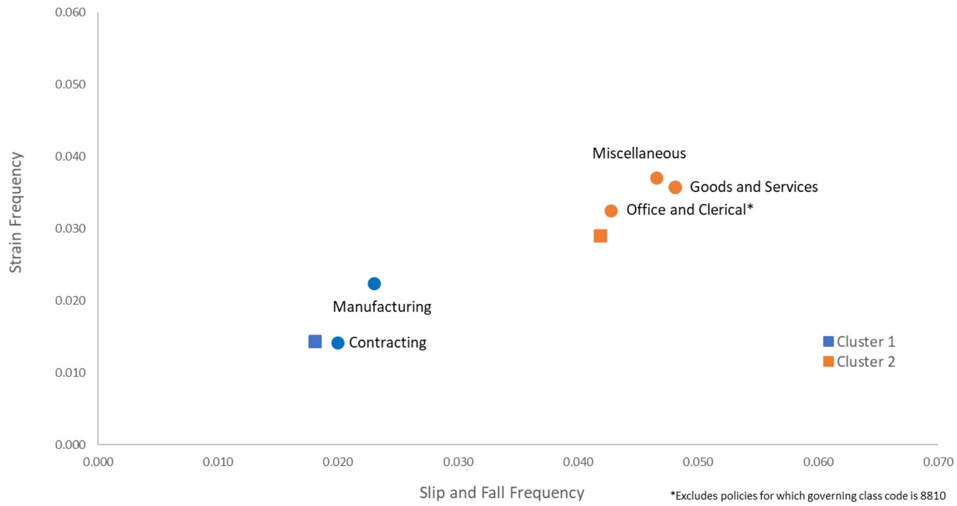


Figure 20 shows cluster means for five cause-of-injury groups. By examining each column, one observes that there is a “well ordering” of these multidimensional vectors. That is, if one cluster has a higher claim frequency for one cause of injury, the other causes of injury also have a higher frequency.

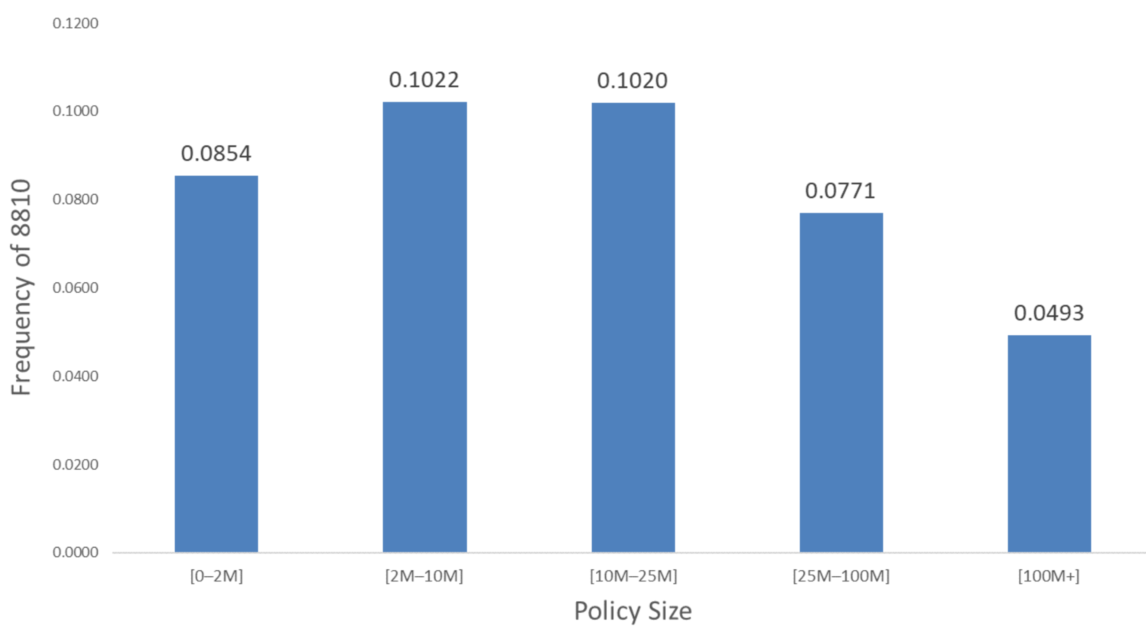
Figure 20: Cluster Means for Five Cause-of-Injury Types

Cluster	Cut	Slip and Fall	Motor Vehicle	Strain	Struck or Injured By
1	0.004	0.018	0.002	0.014	0.006
2	0.007	0.042	0.004	0.029	0.013
3	0.019	0.092	0.009	0.062	0.041

Size of Risk Analysis

Figure 21 shows the Code 8810 claim frequency by size of policy as measured by total payroll. We chose the policy size intervals in such a way that each one contains an equal share of claims. Observe that for policy sizes up to $25 million of total payroll, the frequency is about 0.1[3] or about one claim per $10 million of payroll. As the policy size increases beyond $25 million, the frequency drops considerably.

Figure 21: Code 8810: Claim Frequency by Policy Size
All Governing Class Codes

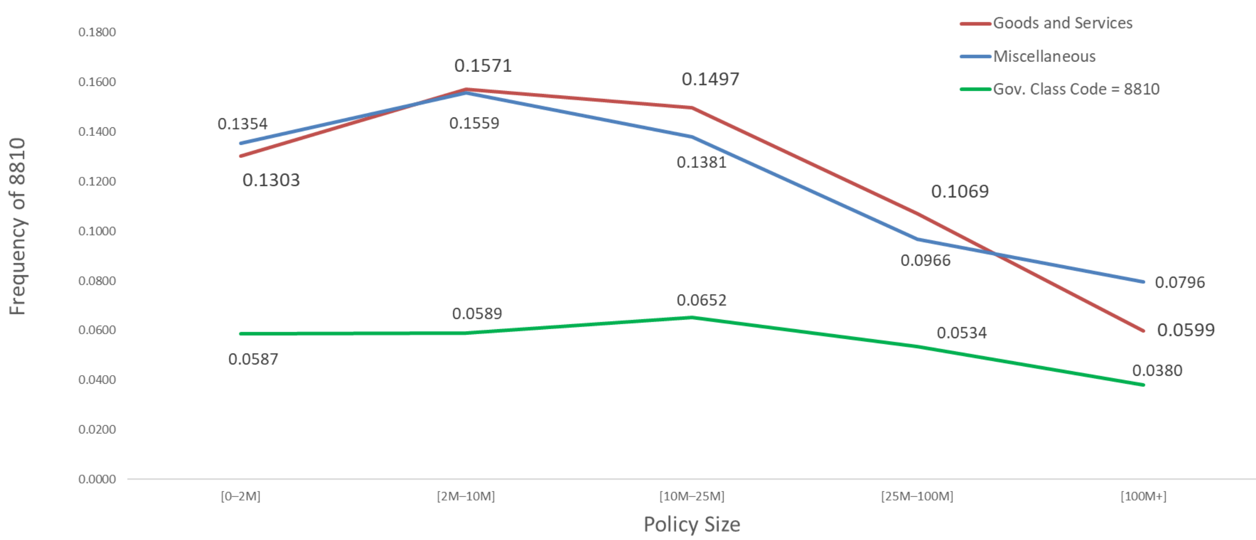


[3] Frequency is defined as the number of claims observed at 1st report (i.e., 18 months maturity) per $1 million of payroll.

Figure 22 compares the claim frequency of Code 8810 for policies that have governing class codes in the Goods and Services and Miscellaneous IGs with the claim frequency of policies whose governing class code is Code 8810. We observed the largest differences for policies below $25 million of payroll. The differences reduce as policy size increases. All differences are statistically significant except for policies below $2 million. It confirms that the earlier-observed trend of higher frequency differentials persists when the governing class code is in the Goods and Services or Miscellaneous IGs, even for large risks.

Figure 22: Code 8810: Claim Frequency by Policy Size

Governing Class Codes in the Goods and Services or Miscellaneous IGs vs. Governing Class Code 8810



CLOSING REMARKS

Claim frequency for O&C workers from policies that have exposures only within the O&C IG is lower than that of O&C workers from policies within other industries. We examined the largest O&C Code 8810 across policies that have various other governing class codes. The analysis shows that the claim frequency varies significantly across governing class codes, with the highest frequency differences observed for governing class codes in the Goods and Services and Miscellaneous IGs. Such differences persist across states as well as across policy sizes.

For years, the most misclassified employees in WC have been within Code 8810. Our claim frequency analysis supports that it is indeed a very heterogeneous classification. Depending on the nature of the business, O&C workers are probably doing other non-O&C tasks, leading to a higher rate of injury than one might otherwise expect. NCCI may explore other O&C classes in the future using similar analytical approaches, such as Code 8742—Salespersons and Collectors, Code 8803—Auditor, Accountant, or Programmer—Traveling, and Code 8871—Clerical Telecommuting Employees.

APPENDIX

Summary of Steps Needed to Perform Cluster Analysis

- Select five causes of injury comprising 80% of all Code 8810 claims:
    - Cut, Puncture, Scrape
    - Slip and Fall
    - Motor Vehicle
    - Strain or Injury By
    - Struck or Injured By
- Select governing class codes comprising 90% of Class Code 8810 payroll. Group the remaining small governing class codes into “catch-all” buckets by IG.
- For each governing class and each catch-all bucket, form a five-dimensional frequency vector based on the five causes of injury above.
- Apply PCA to the original data set to transform the correlated data set into an uncorrelated one. See below in this Appendix for more about PCA.
- PCA analysis also indicates that the optimal number of principal components is two. See below in this Appendix for more details about retaining the optimal number of principal components.
- Apply the cluster analysis on the new, uncorrelated, two-dimensional data.

Frequency, Severity, and Share of Claims by Cause of Injury: Lost-Time Claims Only

Figure A1: Lost-Time Claim Frequency by Cause of Injury
O&C

Cause of Injury	Multi-IG	O&C Only	Relativity	Significant at 5%
Struck or Injured by	0.0038	0.0017	2.21	Yes
Caught in, Under, or Between	0.0007	0.0003	2.17	Yes
Rubbed or Abraded By	0.0001	0.0001	2.07	Yes
Burn	0.0003	0.0002	1.86	Yes
Strain	0.0109	0.0059	1.84	
Cut, Puncture, Scrape	0.0006	0.0003	1.83	
Miscellaneous	0.0029	0.0019	1.51	
Slip and Fall	0.0121	0.0082	1.48	
Striking Against or Stepping On	0.0010	0.0008	1.33	
Motor Vehicle	0.0016	0.0013	1.26	

Figure A2: Lost-Time Claim Severity by Cause of Injury
O&C

Cause of Injury	Multi-IG	O&C Only	Relativity	Significant at 5%
Striking Against or Stepping On	32,198	37,952	0.85	Yes
Burn	31,353	26,560	1.18	
Caught in, Under, or Between	32,863	29,560	1.11	
Rubbed or Abraded By	29,458	27,807	1.06	
Slip and Fall	43,548	43,420	1.00	
Cut, Puncture, Scrape	22,475	24,327	0.92	
Strain	35,372	38,364	0.92	
Miscellaneous	34,556	38,076	0.91	
Motor Vehicle	76,453	89,227	0.86	
Struck or Injured By	36,067	42,810	0.84	

Figure A3: Share of Claims by Cause of Injury for Lost-Time Claims Only
O&C

Cause of Injury	Multi-IG	O&C Only	Significant at 5%
Caught in, Under, or Between	2.2%	1.7%	Yes
Slip and Fall	35.7%	39.8%	Yes
Strain	32.0%	28.7%	Yes
Striking Against or Stepping On	2.9%	3.7%	Yes
Struck or Injured By	11.1%	8.3%	Yes
Burn	0.9%	0.8%	
Cut, Puncture, Scrape	1.6%	1.5%	
Motor Vehicle	4.7%	6.2%	
Rubbed or Abraded By	0.4%	0.3%	
Miscellaneous	8.4%	9.2%	

Frequency and Severity Relativities for Code 8810 for Lost-Time Claims Only

Figure A4: Code 8810 Claim Frequency Relativities by IG of Policy Governing Class

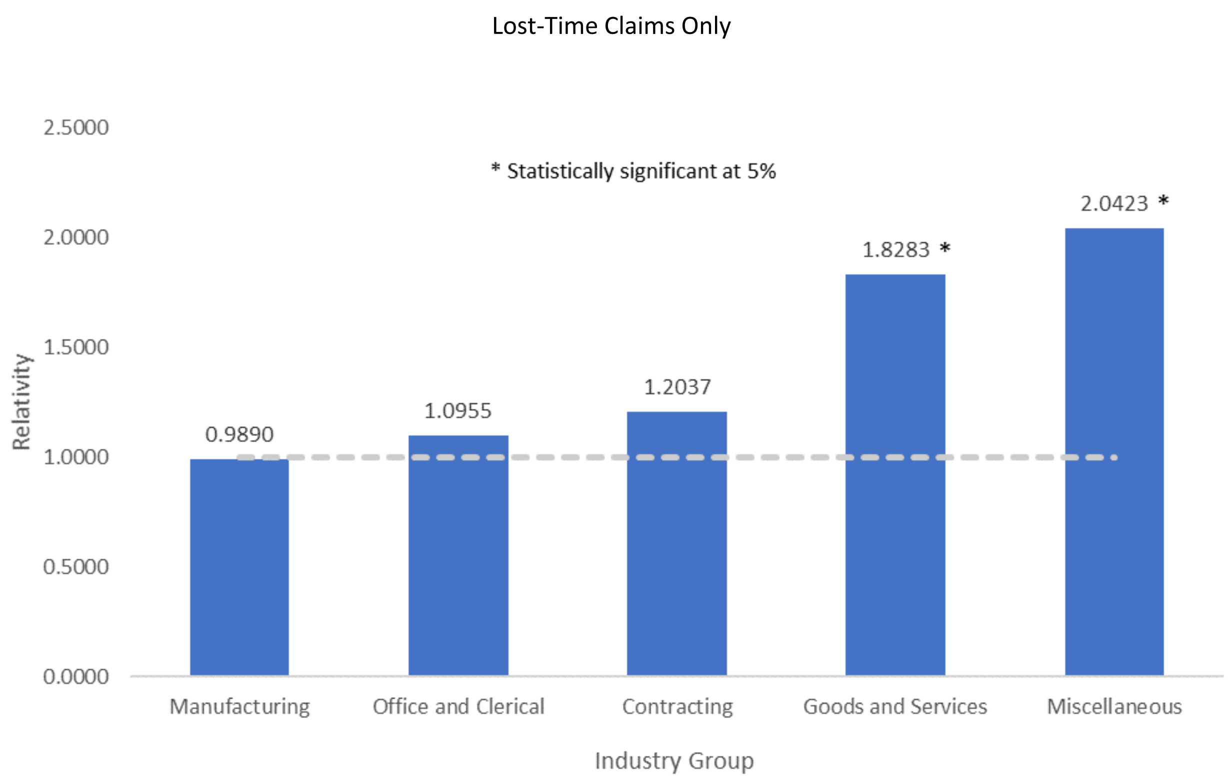
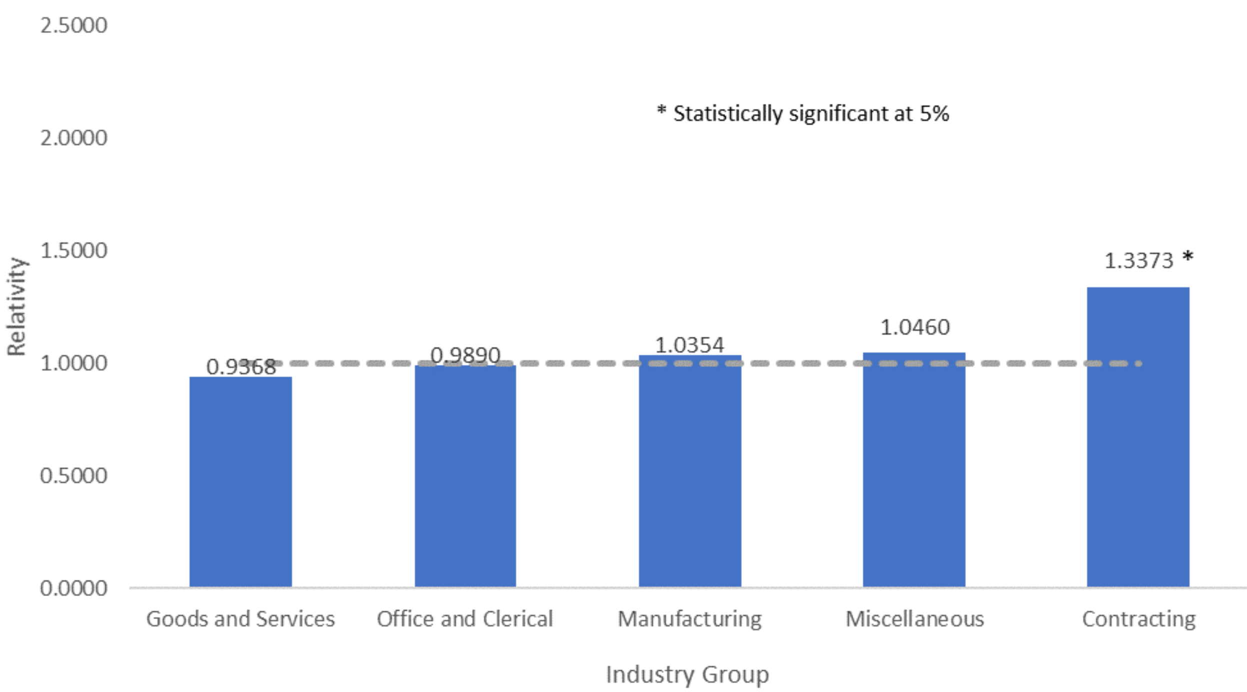


Figure A5: Code 8810 Claim Severity Relativities by IG of Policy Governing Class

Lost-Time Claims Only



Principal Component Analysis (PCA) Detail

PCA provides a linear map ("change of basis" or "change of coordinate system" in linear algebra) that transforms the original vector into a new one. Below is an example illustrating this process.

Assume the following:

- 5-D vector *v*, with each of its components normalized as per footnote 1 on page 8, having coordinates <1, –1, 2, 3, 2>
- PCA provides the following matrix E for the transformation of the original vector

$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_3}$	$\mathbf{e_4}$	$\mathbf{e_5}$
0.41	–0.65	0.22	0.18	0.58
0.48	0.29	–0.39	–0.65	0.33
0.37	0.57	0.72	0.14	0.07
0.48	0.19	-0.51	0.67	–0.15
0.49	–0.36	0.17	–0.28	–0.73

Each column in the above matrix is an eigenvector of the symmetric matrix $X^T X$ where X is the *n*-by-5 data set of governing class codes described in the main body of the brief and T denotes the transpose operator.

Then principal component I is given by the dot product $PC_i = v \cdot e_i$. For example, the first principal component is calculated as:

PC_1 = 1 * 0.41 + (–1) * 0.48 + 2 * 0.37 + 3 * 0.48 + 2 * 0.49 = 3.09

The second principal component is calculated as:

PC_2 = 1 * (–0.65) + (–1) * 0.29 + 2 * 0.57 + 3 * 0.19 + 2 * (–0.36) = 0.05

Furthermore, the matrix $A = E(X^T X)E^T$ is a 5-by-5 diagonal matrix with entries in a_{ii} equal to eigenvalues of E and 0 elsewhere.

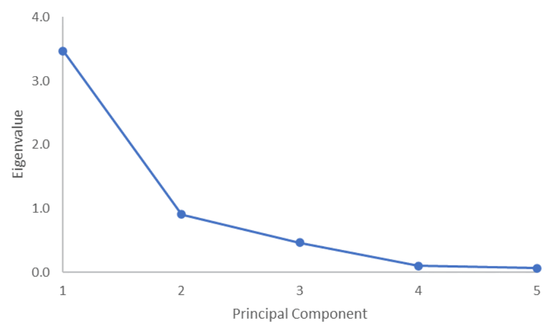
For further details about PCA, refer to [4].

Retaining an Optimal Number of PCA Components

One of the goals of principal components of *n* variables is to represent most of the variation in the data by using *m* new variables where *m* is strictly less than *n* and hopefully much smaller than *n*. There are various methods of retaining an optimal number of PCA components. Some are based on visual inspection of certain diagrams, while others are grounded in probability and statistics. Different methods produce different results.

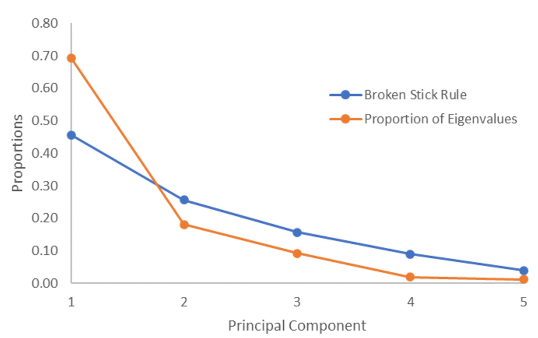
One of the most widely used methods is a scree diagram. If the scree diagram contains a sharp change in the slopes of adjacent segments (an "elbow"), then we choose that point as the number of principal components to retain. Figure A6 shows the scree diagram for the governing class code data set used in the analysis. There exists an abrupt change in the slopes of adjacent segments when the number of principal components is 2. According to the scree diagram, this is the most optimal number of principal components to retain.

Figure A6: Scree Diagram



The broken stick method [1] is one of the probabilistic methods in choosing the number of optimal principal components. This method retains components that explain more variance than would be expected by randomly dividing the variance into *p* parts. If a quantity is randomly divided into *p* parts, the expected proportion of the *k*th largest piece is given by $\frac{\sum_{i=k}^{p} {}^{1}/_{i}}{p}$. Each such expected proportion is compared against the eigenvalue proportions of the $X^T X$ matrix and are calculated as $\frac{EV_k}{\sum_{i=1}^{p} EV_i}$. The observed proportions that are higher than the expected proportions indicate which principal component to retain. Figure A7 compares the eigenvalue proportions against the broken stick rule. According to this method, only one principal component is optimal because the observed proportion of the first eigenvalue is greater than the expected proportion from the broken stick rule.

Figure A7: Broken Stick Diagram



The Kaiser-Guttman test retains eigenvalues that exceed the average eigenvalue. The sum of eigenvalues for a normalized *b*-by-*p* matrix X is *p*, hence the average is 1. The $X^T X$ matrix is 5-by-5, so the sum of eigenvalues is 5 and the average is 1. A more relaxing criterion was proposed by Jolliffe [2], which suggests retaining eigenvalues greater than 0.7.

Eigenvalues of the $X^T X$ matrix are: 3.468, 0.906, 0.462, 0.100, 0.063. The Kaiser-Guttman test suggests retaining only the first principal component because its eigenvalue is greater than 1, while the Kaiser-Guttman-Jolliffe test suggests retaining the first two components because they are greater than 0.7.

In summary, the broken stick and Kaiser-Guttman tests suggest retaining only one principal component, while the scree and Kaiser-Guttman-Jolliffe tests suggest retaining the first two principal components.

For this analysis, we decided to retain the first two principal components because they explain close to 90% of the variance of the original data set.

Impact of Differing Policy Years on Governing Class Code Cluster Assignment

Figure A8 illustrates the cluster assignment for each of the top five governing class codes (ranked by amount of Code 8810 payroll) in each of the three clusters described in the paper when using smaller and different policy years in clustering. Using different policy years can produce a different number of optimal clusters (from four to six). To bring everything to a common denominator, we used five clusters for each iteration. One can think of these clusters as governing class codes with a Code 8810 claim frequency that is:

1. Low
2. Low to medium
3. Medium
4. Medium to high
5. High

Using three different sets of five policy years (PY 2012–2016, PY 2013–2017, and PY 2014–2018) produces remarkable stability in cluster assignments for governing class codes. The PY 2015–2019 group generally shifts the cluster assignments down because PY 2019 shows a significant drop in loss frequency for O&C workers across almost all governing class codes. The 2020 pandemic-related business lockdowns could, in part, have driven this decrease in loss frequency.

Figure A8: Impact of Different Policy Years on Cluster Assignment

Top Five Governing Class Codes in Each of the Three Clusters

Governing Class Code	PY 2012-2016	PY 2013-2017	PY 2014-2018	PY 2015-2019
3681	1	1	1	1
5191	1	1	1	1
8017	1	1	1	1
8601	1	1	1	1
8803	1	1	1	1
8018	2	2	2	1
8380	3	3	3	2
8832	2	2	2	1
8833	3	3	3	2
9015	2	2	2	1
7720	5	4	4	3
8829	5	4	4	3
9014	5	4	4	3
8864	5	4	4	5
9101	5	4	4	3

REFERENCES

[1] Jackson, Donald A., "Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches," *Ecology*, Vol. 74, No. 8, 1993, pp. 2204–2214.

[2] Jolliffe, I. T., "Discarding Variables in a Principal Component Analysis. I: Artificial Data," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 21, No. 2, 1972, pp. 160–173.

[3] Robertson, John P., "NCCI's 2007 Hazard Group Mapping," *Variance Journal*, Vol. 3, Issue 2, 2009, Casualty Actuarial Society, pp.194–213.

[4] Shlens, Jonathon, "A Tutorial on Principal Component Analysis," 2014, Version 3.02.