TOWERS PERRIN

TILLINGHAST

Techniques for Dimension Reduction – Variable Selection with Clustering

CAS Special Interest Seminar on Predictive Modeling

Robert Sanche

October 5, 2006

Contents

- Predictive Variables
  - Economics of Data Storage
  - Sources of Data
  - Redundancy of Variables
- Dimension Reduction
  - Goals of Predictive Modeling
  - Model Generalization
  - Clustering Analysis for Dimensions Reduction
- Variable Clustering
  - Description of Variable Clustering
  - Selection of the Cluster Representative
  - Example of Variable Clustering
- Conclusion and Benefits of Variable Clustering

Economics of Data Storage

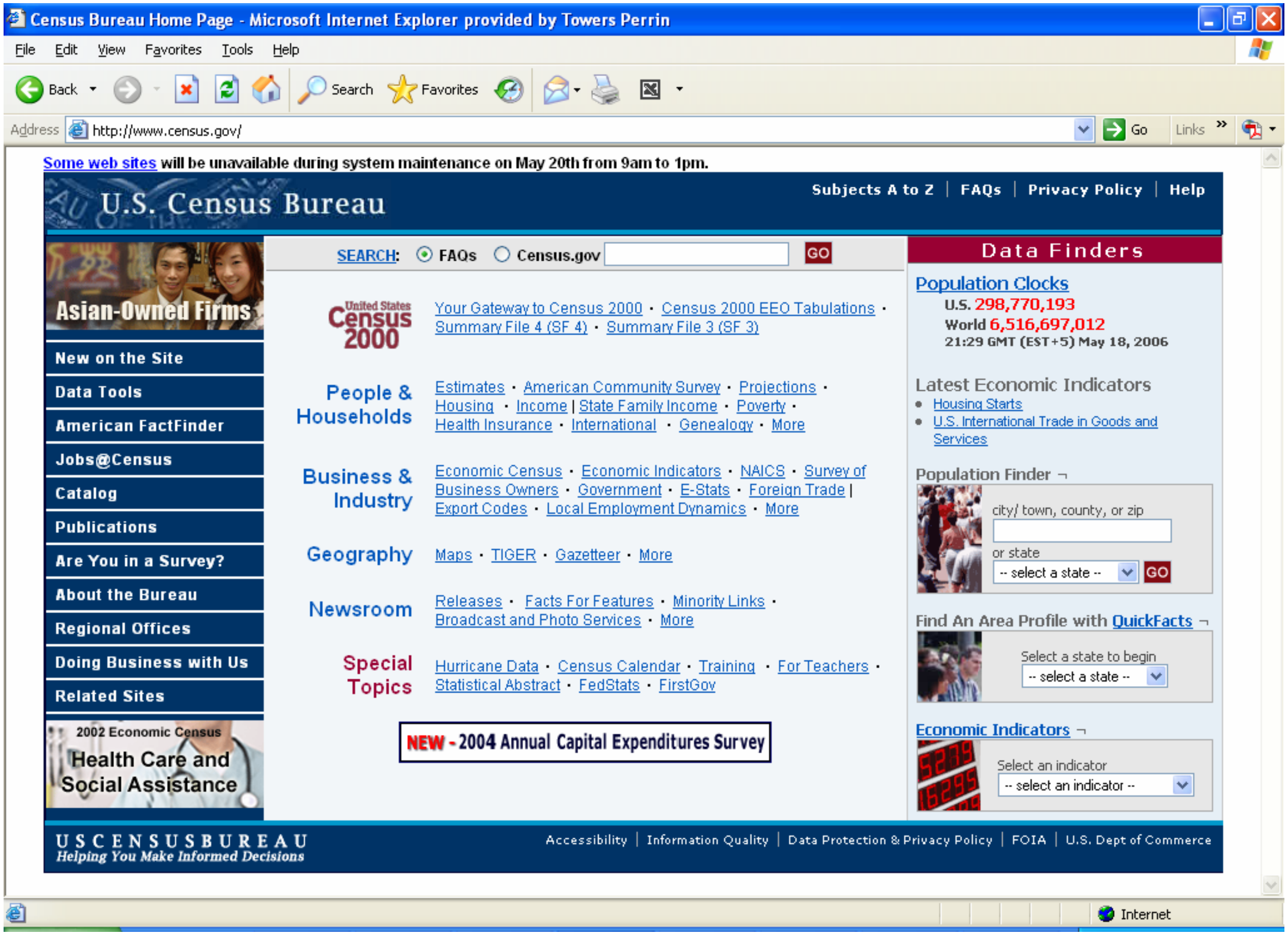
"In 1956, IBM sold its first magnetic disk system, RAMAC (Random Access Method of Accounting and Control). It used 50 24-inch metal disks, with 100 tracks per side. It could store 5 megabytes of data and cost $10,000 per megabyte. (As of 2005, disk storage costs less than $1 per gigabyte)."
http://en.wikipedia.org/wiki/History_of_computing_hardware

- 1 gigabyte = 130 numeric characteristics
  - for 1 million policies
  - for $1.00

Sources of Data

- New data sources
  - Data warehousing (coverage and claims)
  - External sources
    - — Geo-demographics
    - — Meteorological
    - — Policyholder, household, business owner, company information or agent
  - Other

External Data (Census)



Census Bureau Home Page - Microsoft Internet Explorer provided by Towers Perrin
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
http://www.census.gov/
Go
Links
Internet

Some web sites will be unavailable during system maintenance on May 20th from 9am to 1pm.

U.S. Census Bureau

Subjects A to Z | FAQs | Privacy Policy | Help

SEARCH: ● FAQs ○ Census.gov GO

Asian-Owned Firms

- New on the Site
- Data Tools
- American FactFinder
- Jobs@Census
- Catalog
- Publications
- Are You in a Survey?
- About the Bureau
- Regional Offices
- Doing Business with Us
- Related Sites

2002 Economic Census
Health Care and Social Assistance

United States Census 2000 — Your Gateway to Census 2000 • Census 2000 EEO Tabulations • Summary File 4 (SF 4) • Summary File 3 (SF 3)

People & Households — Estimates • American Community Survey • Projections • Housing • Income | State Family Income • Poverty • Health Insurance • International • Genealogy • More

Business & Industry — Economic Census • Economic Indicators • NAICS • Survey of Business Owners • Government • E-Stats • Foreign Trade | Export Codes • Local Employment Dynamics • More

Geography — Maps • TIGER • Gazetteer • More

Newsroom — Releases • Facts For Features • Minority Links • Broadcast and Photo Services • More

Special Topics — Hurricane Data • Census Calendar • Training • For Teachers • Statistical Abstract • FedStats • FirstGov

NEW - 2004 Annual Capital Expenditures Survey

Data Finders

Population Clocks
U.S. 298,770,193
World 6,516,697,012
21:29 GMT (EST+5) May 18, 2006

Latest Economic Indicators
- Housing Starts
- U.S. International Trade in Goods and Services

Population Finder
city/ town, county, or zip
or state
-- select a state -- GO

Find An Area Profile with QuickFacts
Select a state to begin
-- select a state --

Economic Indicators
Select an indicator
-- select an indicator --

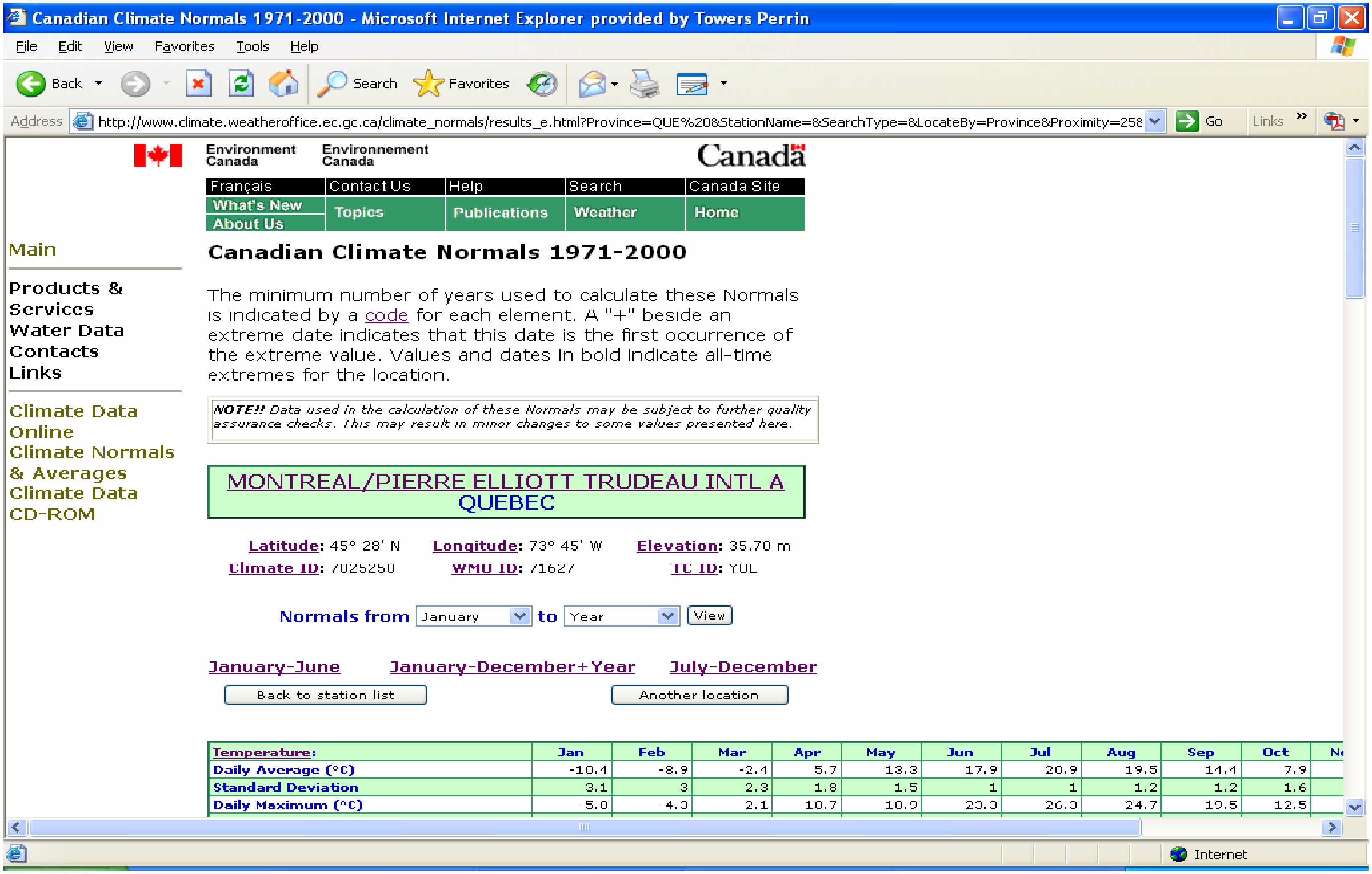
U S C E N S U S B U R E A U
Helping You Make Informed Decisions

Accessibility | Information Quality | Data Protection & Privacy Policy | FOIA | U.S. Dept of Commerce

Census (Geo-demographics)

- Population
  - Average household size
  - Median household size
  - Population density
  - Proportion of household with more than 4
  - Etc.

Meteorological (Environmental Canada)



Canadian Climate Normals 1971-2000

The minimum number of years used to calculate these Normals is indicated by a code for each element. A "+" beside an extreme date indicates that this date is the first occurrence of the extreme value. Values and dates in bold indicate all-time extremes for the location.

NOTE!! Data used in the calculation of these Normals may be subject to further quality assurance checks. This may result in minor changes to some values presented here.

MONTREAL/PIERRE ELLIOTT TRUDEAU INTL A
QUEBEC

Latitude: 45° 28' N Longitude: 73° 45' W Elevation: 35.70 m
Climate ID: 7025250 WMO ID: 71627 TC ID: YUL

Normals from January to Year View

January-June January-December+Year July-December

Back to station list Another location

Temperature:	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	No
Daily Average (°C)	-10.4	-8.9	-2.4	5.7	13.3	17.9	20.9	19.5	14.4	7.9	
Standard Deviation	3.1	3	2.3	1.8	1.5	1	1	1.2	1.2	1.6	
Daily Maximum (°C)	-5.8	-4.3	2.1	10.7	18.9	23.3	26.3	24.7	19.5	12.5	

Meteorological (Temperature)

Temperature	**Days with Minimum Temperature**
Daily Average (°C)	> 0 °C
Standard Deviation	<= 2 °C
Daily Maximum (°C)	<= 0 °C
Daily Minimum (°C)	< -2 °C
	< -10 °C
Degree Days	< -20 °C
Above 24 °C	< - 30 °C
Above 18 °C	
Above 15 °C	**Days with Maximum Temperature**
Above 10 °C	<= 0 °C
Above 5 °C	> 0 °C
Above 0 °C	> 10 °C
Below 0 °C	> 20 °C
Below 5 °C	> 30 °C
Below 10 °C	> 35 °C
Below 15 °C	
Below 18 °C	

Meteorological (Precipitation)

Precipitation	**Days with Rainfall**
Rainfall (mm)	>= 0.2 mm
Snowfall (cm)	>= 5 mm
Precipitation (mm)	>= 10 mm
Average Snow Depth (cm)	>= 25 mm
Median Snow Depth (cm)	
Snow Depth at Month-end (cm)	**Days With Snowfall**
	>= 0.2 cm
Days with Precipitation	>= 5 cm
>= 0.2 mm	>= 10 cm
>= 5 mm	>= 25 cm
>= 10 mm	
>= 25 mm	**Days with Snow Depth**
	>= 1 cm
	>= 5 cm
	>= 10
	>= 20

Redundancy of Variables

- External sources of data are highly redundant
- Note that the data is almost exclusively numeric
  - This fact is primordial in order to use variable clustering

Goals of Predictive Modeling

- Predictive model
  - **$Y = \alpha_1 X_1 + \ldots + \alpha_n X_n + \beta$**
    - **n is universe of all available predictors**

- Goal of predictive modeling
  - Obtain coefficients for α’s and β
- Additional goal
  - Predictive of future results
  - Model generalizes well over time

Model Generalization

- As the number of variables increases and the model complexity increases, the potential of over-fitting the input data increases
- Dimensions reduction
  - Clustering (K-Means)
    - Rows
  - variable clustering
    - Columns
    - Alternatives (Factor, PCA, One-way)

Clustering Analysis for Dimensions Reduction

- "**Cluster Analysis** is a set of methods for constructing a sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual"

  B.S. Everitt , *The Cambridge Dictionary of Statistics*, 1998

- Divide set of data (variables) into groups of similar characteristics
- Unsupervised learning technique
- Useful only when there is **redundancy** in the data

Description of Variable Clustering

- **Variable clustering** divides a set of numeric variables into clusters.
- A large set of variables can be replaced by a single member (cluster representative).
- Reduce the number of variables
  - More difficult to identify irrelevant variables than redundant variables

- **$Y = \alpha_1 X_1 + \ldots + \alpha_m X_m + \beta$**
  - **where $m<n$**

Selection of the Cluster Representative

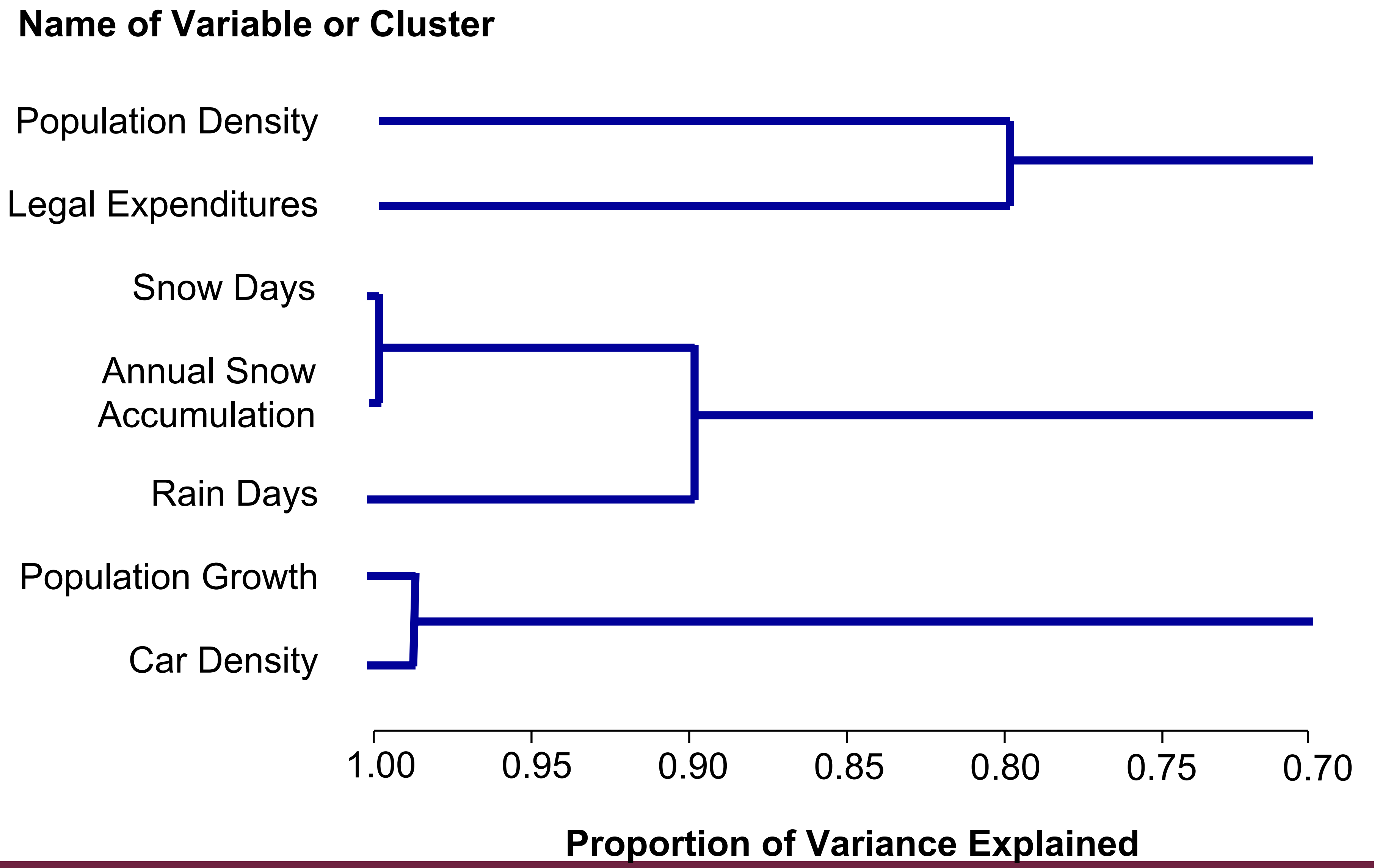
$$1\text{-}R^2_{ratio} = (1\text{-}R^2_{own})/(1\text{-}R^2_{nearest})$$

- Intuitively, we want the cluster representative to be as closely correlated to its own cluster ($R^2_{own}\rightarrow 1$) and as uncorrelated to the nearest cluster ($R^2_{nearest}\rightarrow 0$).
- Therefore, the optimal representative of a cluster is a variable where $1\text{-}R^2$ ratio tends to zero

Example of Variable Clustering

3 CLUSTERS		R-SQUARED WITH		1-R^2 Ratio
Cluster	Variable	Own Cluster	Next Closest	
Cluster 1	Rain Days	0.5995	0.0426	0.4183
	Snow Days	0.8976	0.0317	0.1095
	Annual Snow	0.8940	0.0314	0.1095
Cluster 2	Population Density	0.9804	0.0228	0.0201
	Car Density	0.9804	0.0113	0.0199
Cluster 3	Population Growth	0.6459	0.0911	0.3896
	Legal Expenditures	0.6459	0.0013	0.3546

Clusters of Variables



Conclusion and Benefits of Variable Clustering

- Variable clustering reduces the amount of variables available for predictive modeling (GLM, etc.)
- The predictive modeling process using variable clustering
  - Produces a model that generalize well over time
  - Increases interpretability of the results
  - Reduces time spend on variables selection