

Disease Prevention to Reduce New Hampshire Healthcare Claims and Costs:

A Data Mining Approach

Potential
of One

Power
of
AI

Disease Prevention to Reduce New Hampshire Healthcare Claims and Costs: A Data Mining Approach

Rakesh Karn, Rom Khattri, Pradeep Podila, Linda Schumacher

Oklahoma State University, Stillwater, OK

Abstract

The healthcare industry in the United States is going through a paradigm shift, moving away from its focus on treating diseases and toward promoting health, wellness, and preventive public health programs, so that both the individuals and the government can maintain a healthy bottom line. The high-level business problem is to reduce the expected medical costs and number of medical services required by the people of New Hampshire by implementing successful disease prevention programs. The objective is to identify which among the six prevention programs will successfully improve the health of the residents of New Hampshire over nine future years (2012–2020). The business scenario of the case is to identify the preventive programs that are most effective in reducing the costs in New Hampshire and to invest the money in those programs so that the overall health-care overhead costs can be reduced or controlled. The effectiveness of implementing the preventive programs was evaluated using SAS® Enterprise Guide® 5.1 and SAS® Enterprise Miner™ 12.1 Time series analysis, in particular, forecasting, is used to project the future health-care services and costs for the years from 2012 to 2020.

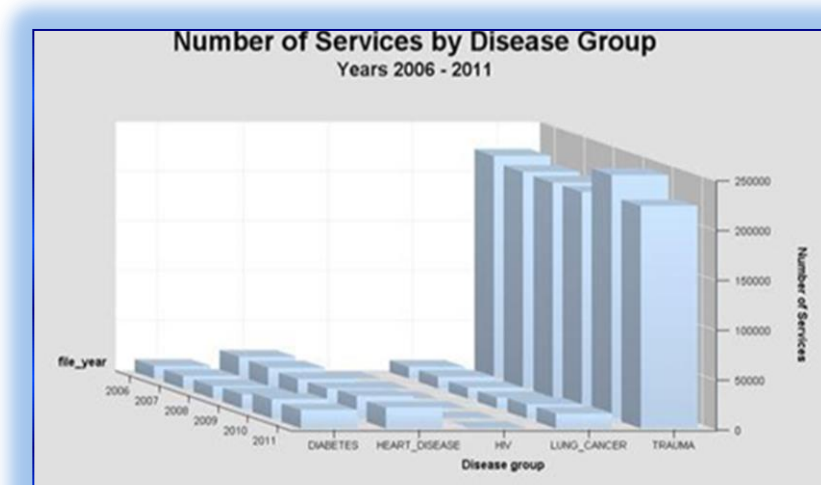
Objective

The objective of this analysis was to forecast the reduction in the number of medical services and costs associated with the implementation of five possible disease prevention programs in the state of New Hampshire. The disease groups under consideration were diabetes, heart disease, HIV, lung cancer and trauma. The results of implementing the prevention programs individually or concurrently were evaluated.

Data

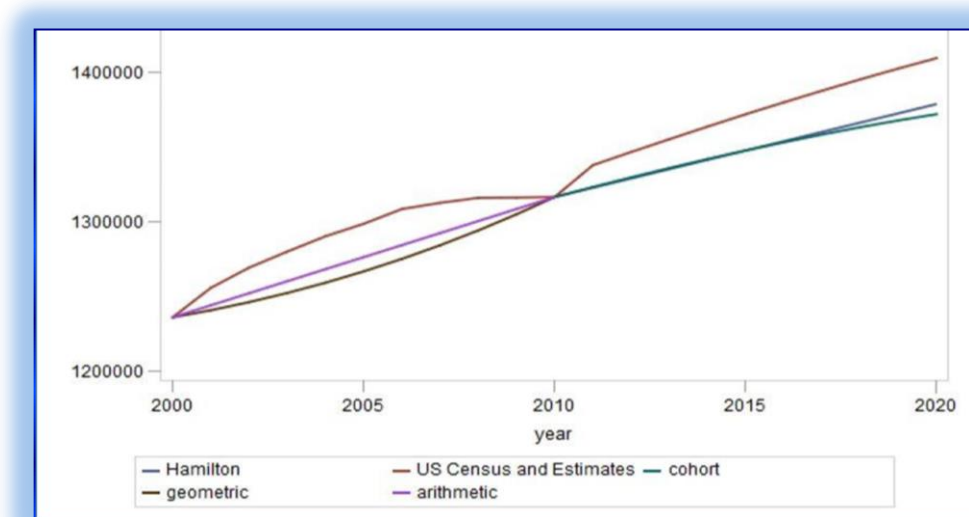
The data were obtained from the Analytics 2013 Data Shootout committee. Three spreadsheet files contained NH population information: 2000 US Census figures, 2010 US Census figures and NH birth and death rates. Medical and pharmacy claims information for the years 2006–2010 were provided in SAS data sets. Relationships between the data files were identified by the service id, diagnostic code, county, age range, and gender. The data were manipulated to obtain a dataset of services related to the five disease groups.

The dataset contained one record with the total medical cost per service. A second record with the total pharmacy costs was present if pharmacy costs were found related to the medical service entry. The number of trauma services was far larger than any other. The median cost per service was highest for lung cancer and heart disease.

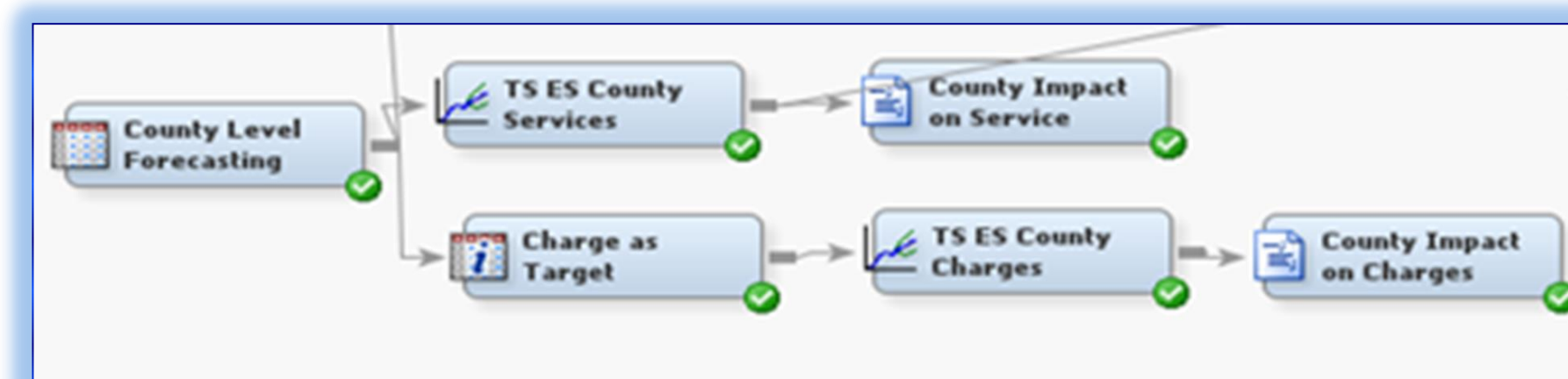


Methods

The future medical and pharmacy claims were dependent on population changes in New Hampshire. Intercensal population estimates were made using arithmetic and geometric methods. Population projections for the years 2011–2020 were made by Hamilton-Perry and cohort survival methods. After comparing the projections with US Census bureau estimates the arithmetic interpolation method for intercensal years and the cohort component method for future predictions were selected.

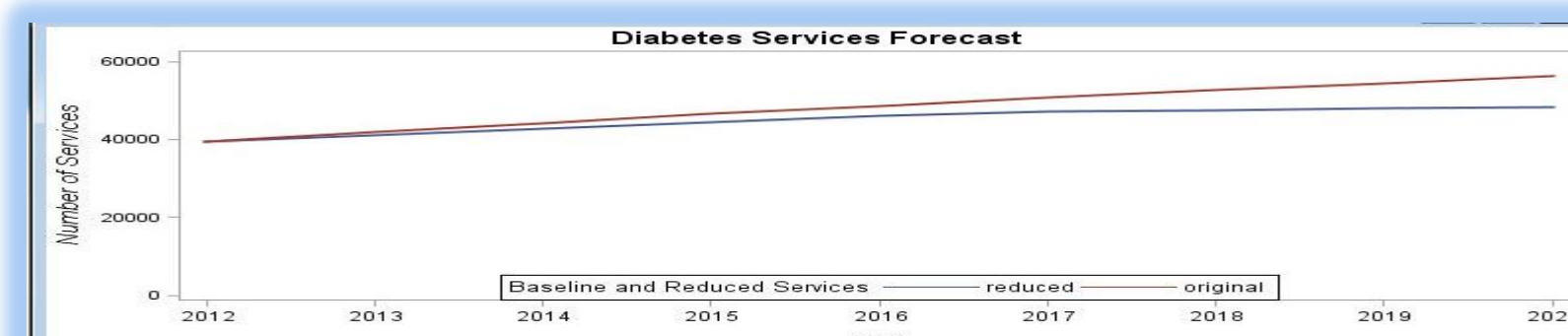


Time series analysis was used to project the future health care services and costs. The population estimates and claims data were used to calculate the incidence rates by disease group, age range, county and claim type. SAS® Enterprise Miner™ 12.1 time series nodes were used to forecast the future incidence rates, using service_per_capita, as the target variable. The candidate exponential smoothing techniques were double, linear, and damped trend smoothing. A second time series node using cost_per_service as the target variable forecast costs using the same smoothing candidate techniques.



Results

The estimates of the future costs by disease group and cohort were calculated using the most likely forecast figures. The improvement impact of the six prevention programs were given as the percent reduction in propensity. These percent reductions and most likely forecast number of services were used to calculate the reductions in the number of future services.



Running the Programs Concurrently
Savings in Services and Costs

		services	services_red	ServicesSaved	baseline_charges	concurrent_charges	CostSaved
rectype							
MC	Sum	1,933,920	1,836,023	97,897	\$15,370,993	\$14,873,164	\$497,829
PC	Sum	244,728	232,116	12,613	\$1,908,938	\$1,834,656	\$74,282
Sum		2,178,649	2,068,139	110,510	\$17,279,931	\$16,707,819	\$572,111

Conclusions

- Because of synergy or interaction effects, implementing all programs concurrently provided the greatest cost savings.
- The health prevention programs under consideration will benefit the people of New Hampshire. It is recommended that all programs are implemented and run concurrently. The minimum anticipated savings in cost is approximately \$572,111 or 3.3% of the expected baseline cost of \$17,279,931. This amount should be used as the initial funding for the prevention programs being run concurrently.

References

- American Public Health Association, June 2012. The Prevention and Public Health Fund: A Critical Investment in Our Nation's Physical and Fiscal Health.
- Brocklebank J, Lee T, Leonard M. Forecasting Cross-Sectional Time Series: A Data Mining Approach Using Enterprise Miner Software.
- CDC. National Prevention Strategy: America's Plan for Better Health and Wellness.
- Census Bureau (October 2012). Methodology for the Intercensal Population and Housing Unit Estimates: 2000 to 2010.
- Department of Health and Human Services. Fiscal Year 2012. Justification of Estimates for Appropriation Committees.
- Henry B, Russo P. (2009). Health Reform: Time for a Paradigm Shift. Views from the Field. Grantmakers in Health.
- Improve Predictive Models - And Decision Making - By Combining Time Series Analysis And Data Mining. SAS.com magazine.
- Public Service of New Hampshire. Fact Book: Demographics (2011). Compiled by: The Northeast Utilities System. August 2011.
- Rey TD, Kaul J. Using Data Mining in Forecasting Problems (Paper 085-2013). The Dow Chemical Company; Chip Wells, SAS Institute Inc., Tata Consultancy Services.
- Schubert S, Lee T. Time Series Data Mining with SAS® Enterprise Miner™. SAS Global Forum 2011.
- Scott RD, Solomon SL, McGowan JE. Applying Economic Principles to Health Care. CDC and Emory University, Atlanta, GA.
- Tang G. Text and Data Mining to Investigate Expenditures on Prescribed Medicines. University of Louisville, KY.

Data Exploration and Preparation

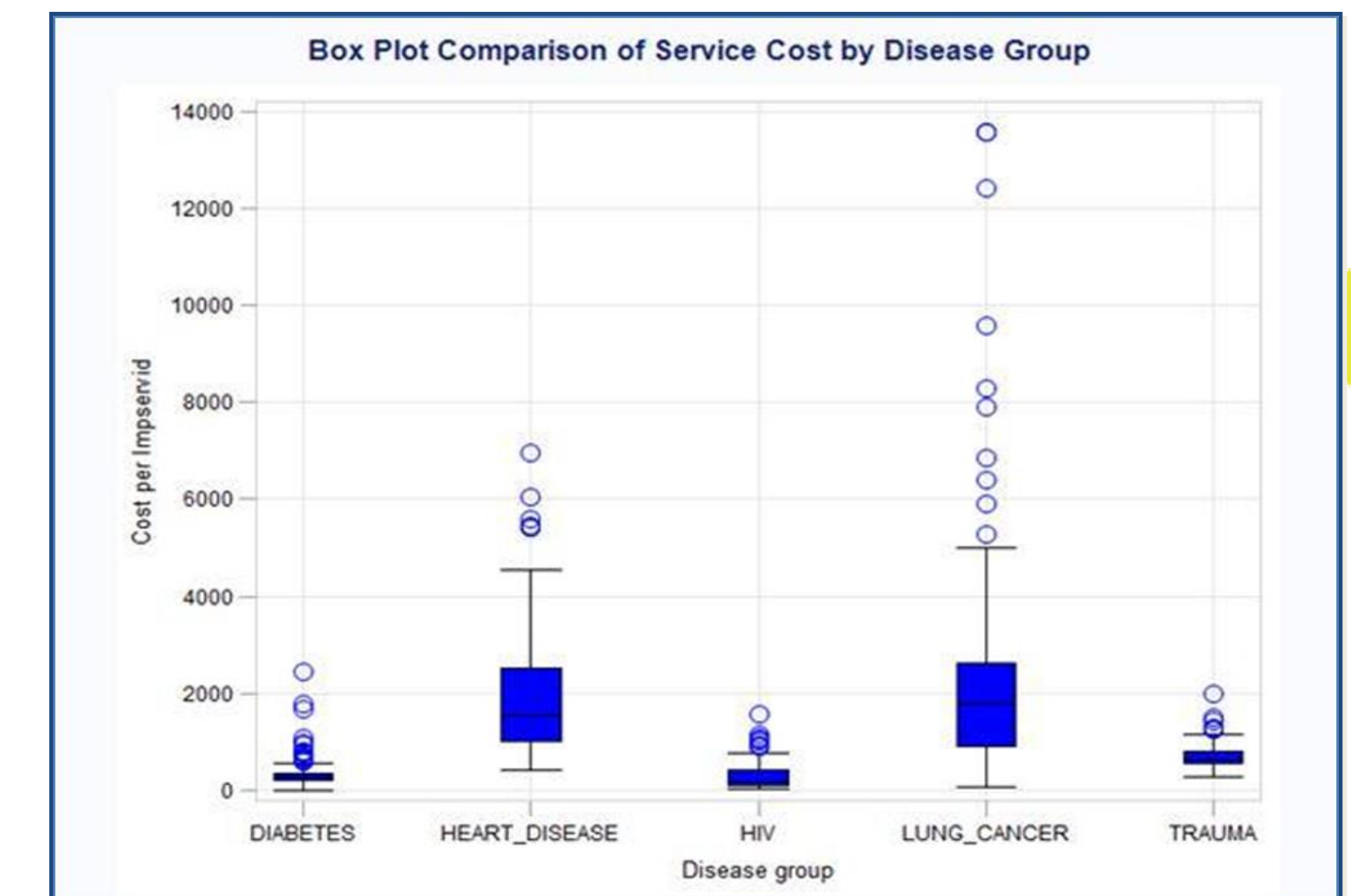
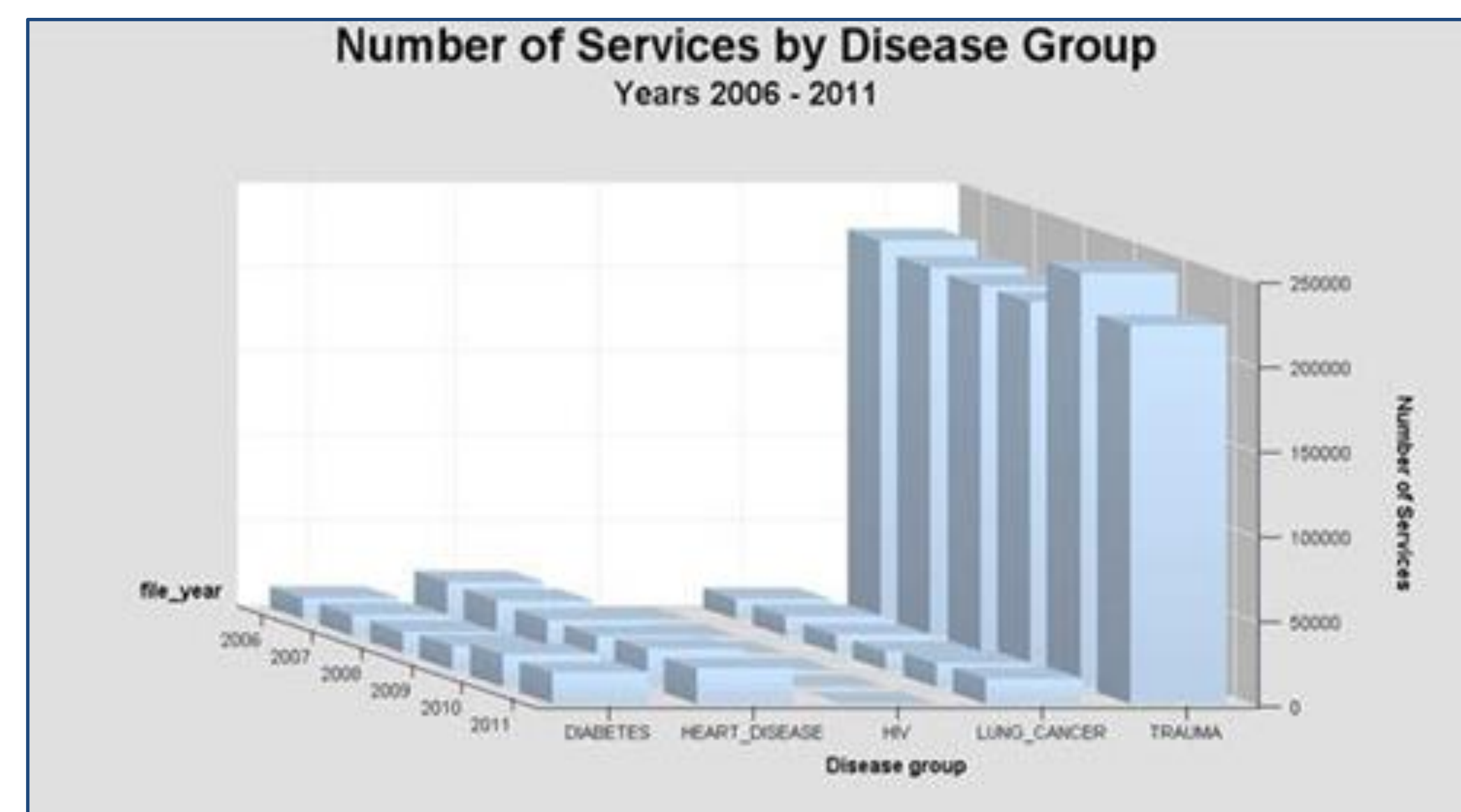
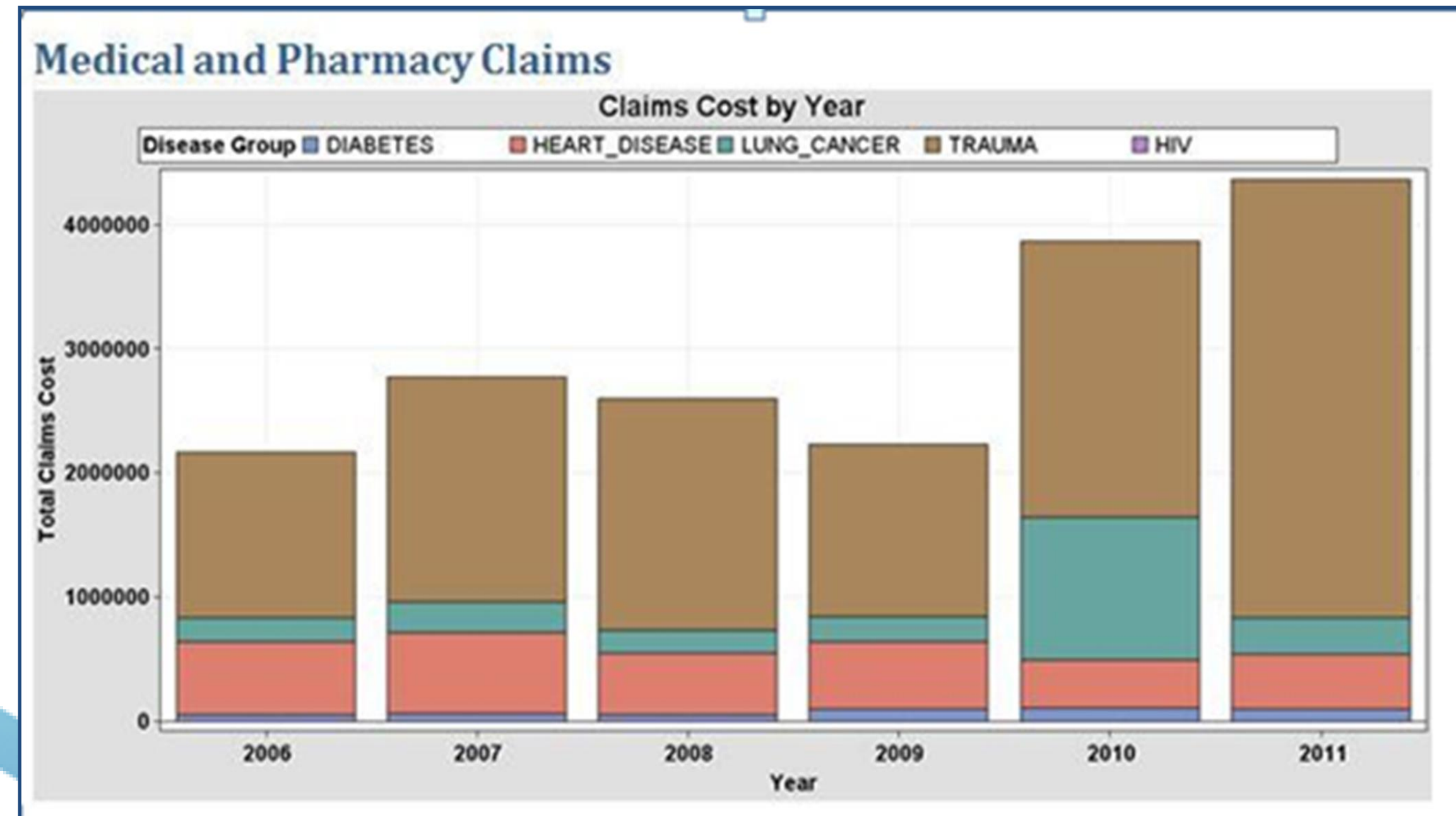
SAS dataset	Records	Services	Common Service ID	Records Common Service ID
Medical Claims	3,906,287	1,500,186	185,270	-----
Pharmaceutical claims	1,927,153	1,791,002		266,911

- Identify prevention programs that will improve the health of the residents of New Hampshire.
 - BMI reduction
 - Stop smoking
 - Cholesterol control
 - Safety program
 - Drug rehab
 - Safe sex programs
- Improvement measured by forecasted reduction in cost and services

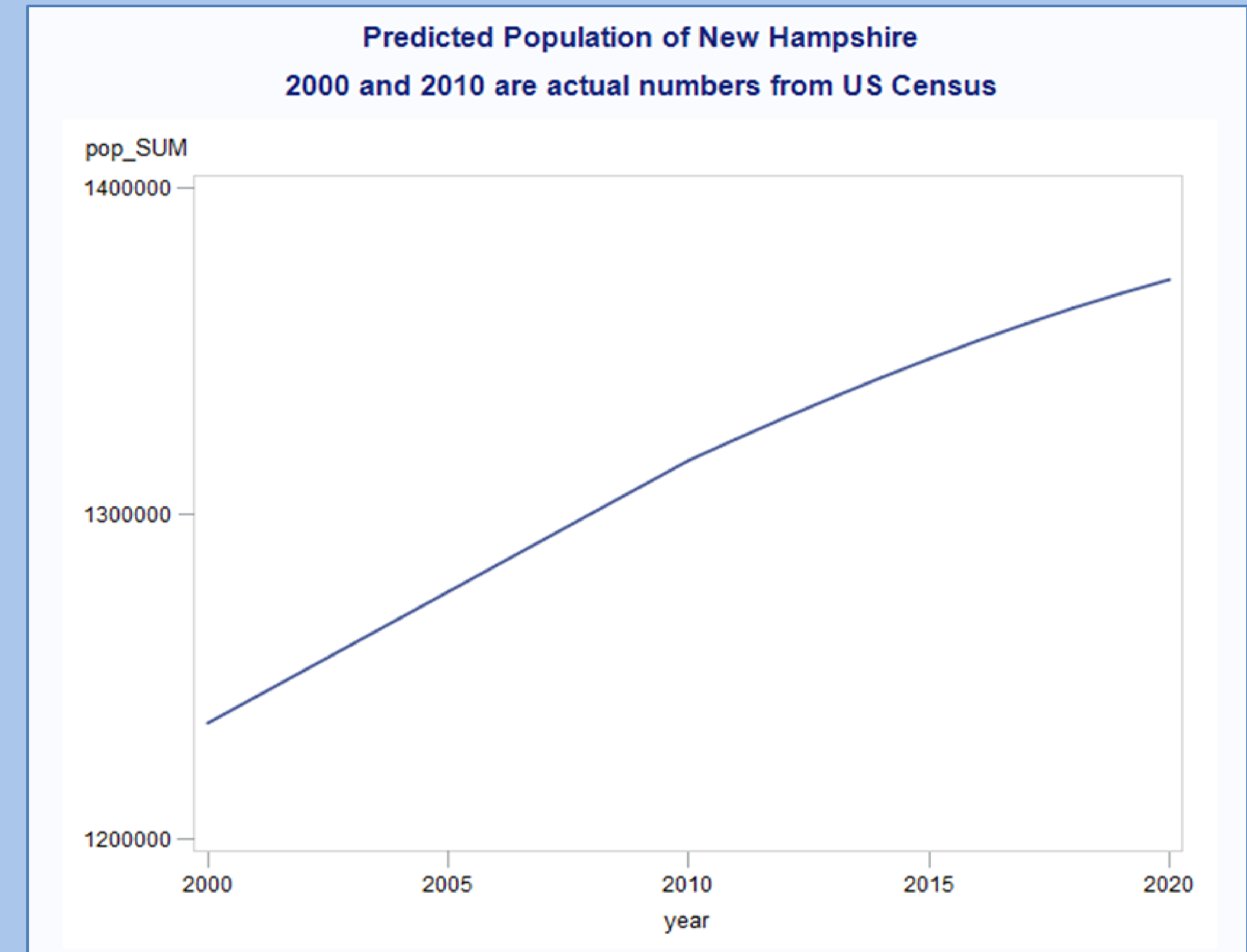
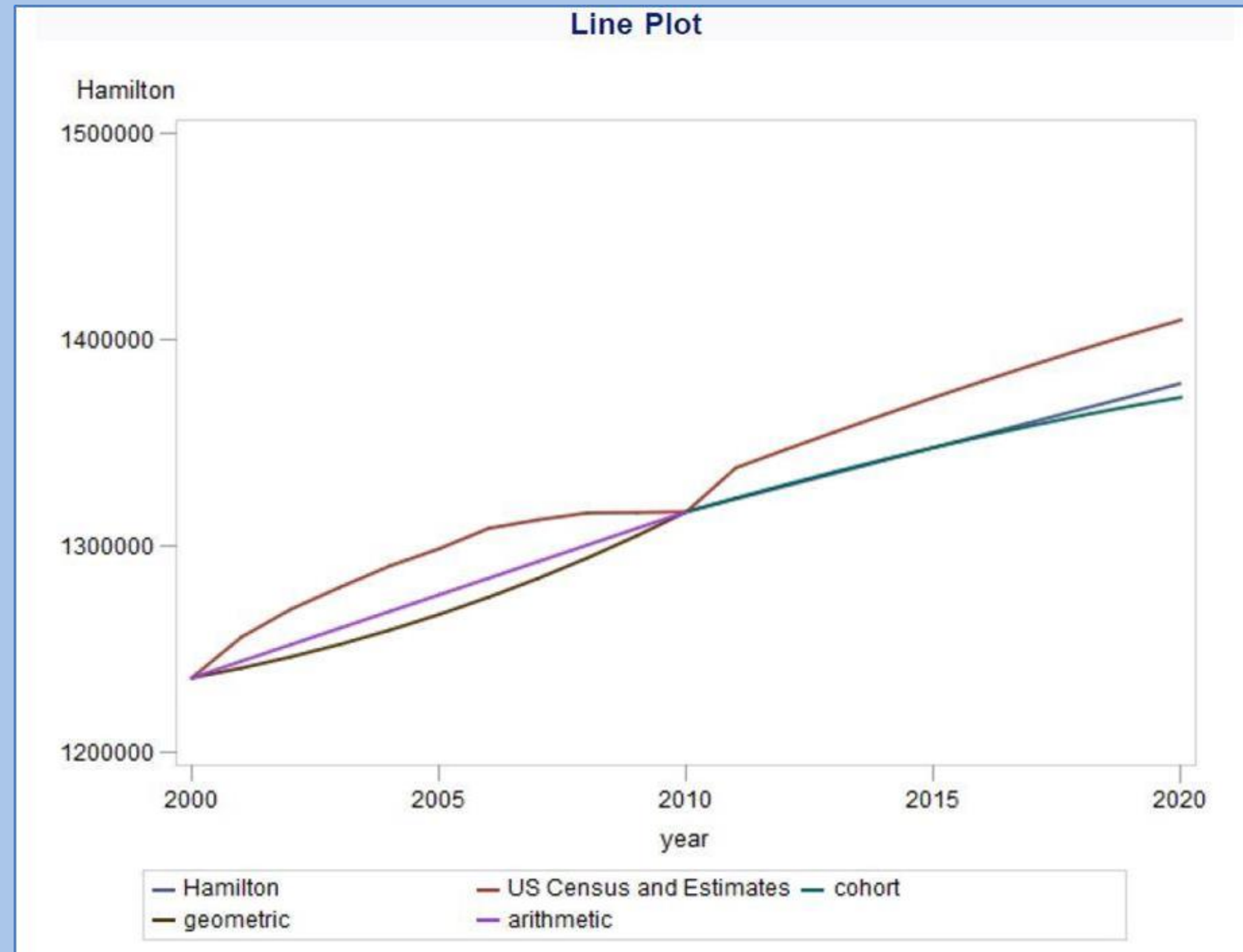
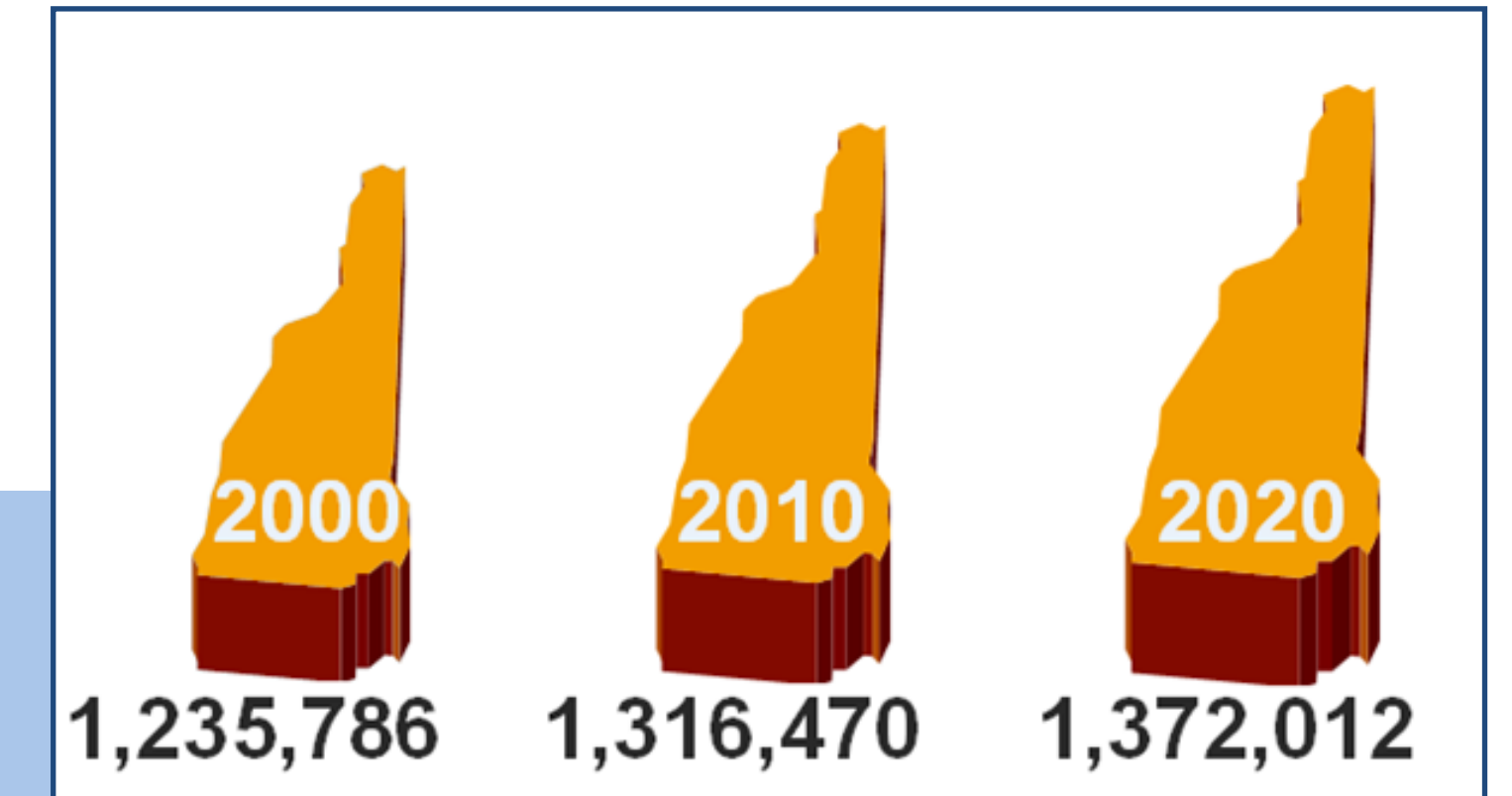
- Integrate
 - Merge datasets by key variables
 - Select medical claims by ICD-9 codes
 - Match medical & pharma claims by service ID
- Pharmacy Refills for chronic conditions
- Filter Negative Cost
- Cutoff Cost at Extreme Percentiles

Refills - Pharmaceutical Claims Only

- PROC SQL algorithm to find most frequent disease group per first word of drug name
- Heart Disease Pharma Claims
- Diabetes Pharma Claims
- Caveat – No county identification



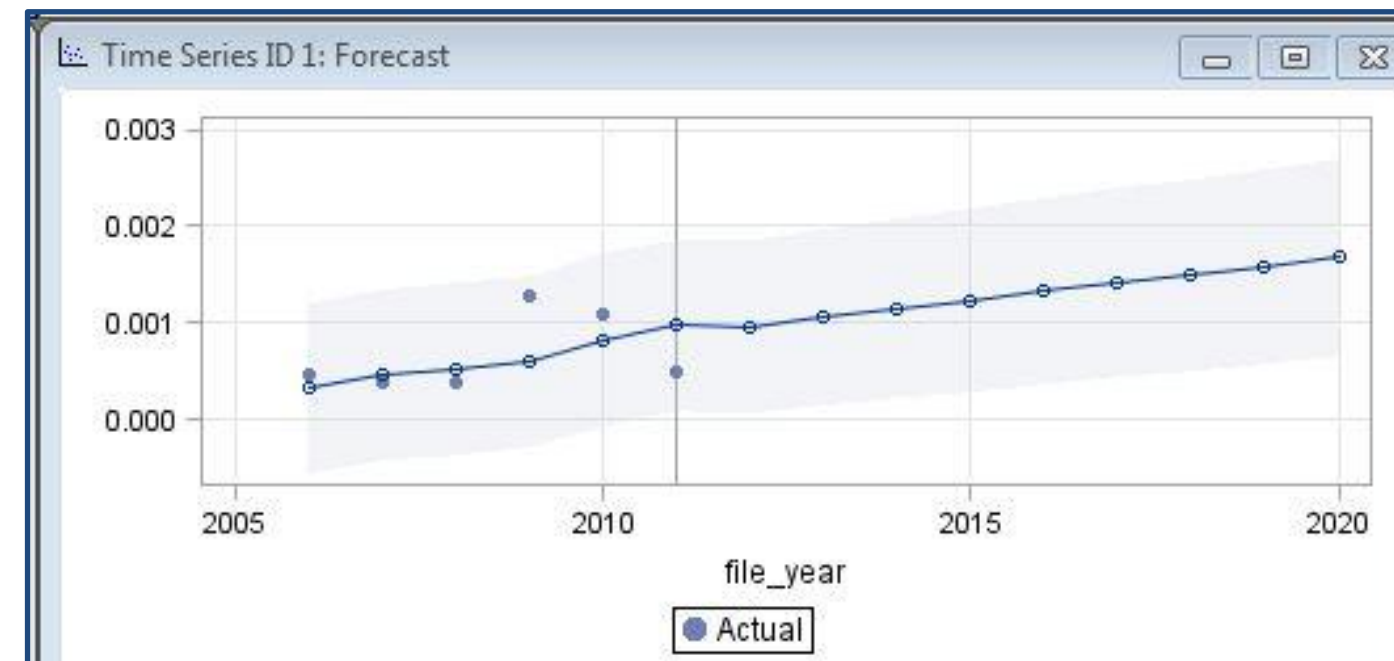
Population Estimation



- 2000–2010 Intercensal: **Arithmetic**, Geometric
- 2011–2020 Projections: **Cohort Survival**, Hamilton Perry



Time Series Exponential Smoothing



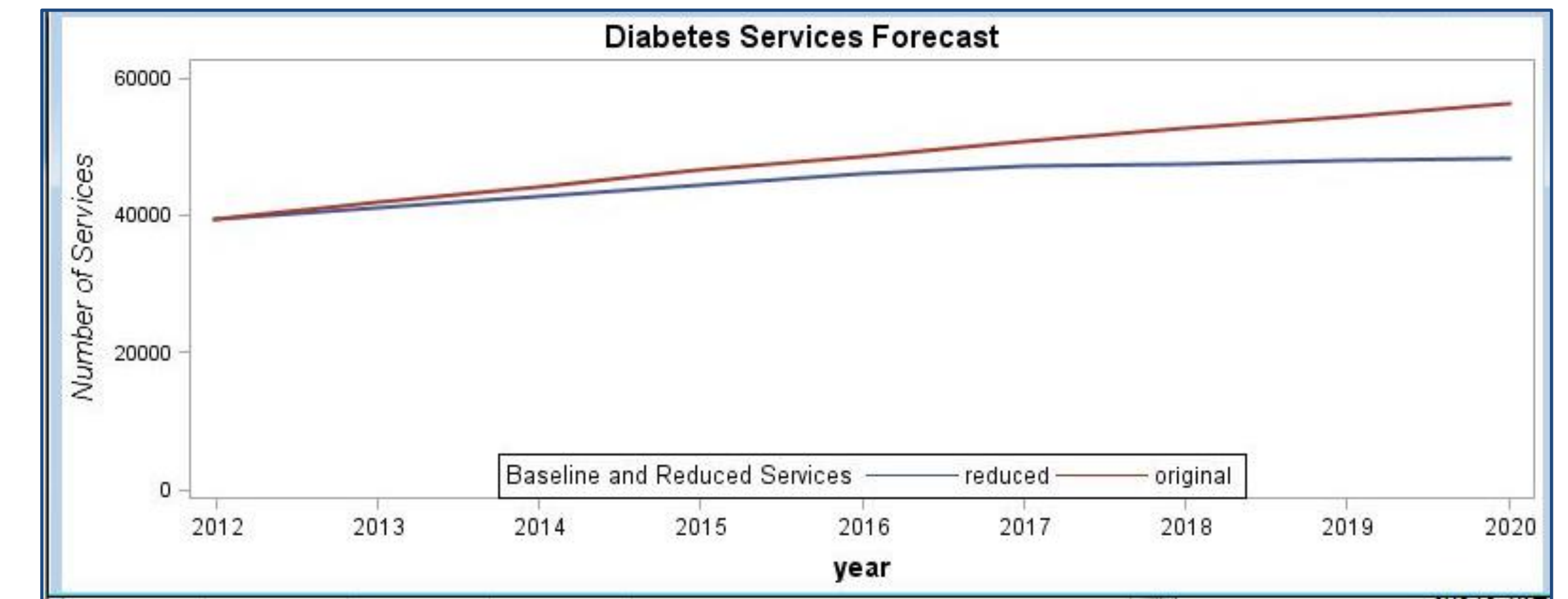
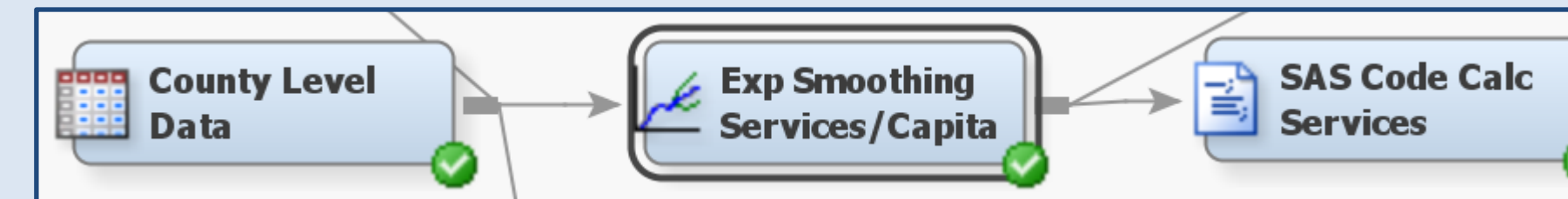
TSID 1: 0<15 yrs Diabetes Medical Claims Belknap County

Time Series ID	file_year	County_FIPS	age_range	Forecasting Model	Actual Values	Predicted Values	Disease group	rectype	Variable Name
1	2006	330010<15		LINEAR	.0004827	.0003379	DIABETES	MC	service_ca...
1	2007	330010<15		LINEAR	.0003892	.0004573	DIABETES	MC	service_ca...
1	2008	330010<15		LINEAR	.0003923	.0005328	DIABETES	MC	service_ca...
1	2009	330010<15		LINEAR	.0001285	.0005934	DIABETES	MC	service_ca...
1	2010	330010<15		LINEAR	.0001097	.0008256	DIABETES	MC	service_ca...
1	2011	330010<15		LINEAR	.0005089	.0009711	DIABETES	MC	service_ca...
1	2012	330010<15		LINEAR		.0009655	DIABETES	MC	service_ca...
1	2013	330010<15		LINEAR		.0001055	DIABETES	MC	service_ca...
1	2014	330010<15		LINEAR		.0001145	DIABETES	MC	service_ca...
1	2015	330010<15		LINEAR		.0001234	DIABETES	MC	service_ca...
1	2016	330010<15		LINEAR		.0001324	DIABETES	MC	service_ca...
1	2017	330010<15		LINEAR		.0001414	DIABETES	MC	service_ca...
1	2018	330010<15		LINEAR		.0001503	DIABETES	MC	service_ca...
1	2019	330010<15		LINEAR		.0001593	DIABETES	MC	service_ca...
1	2020	330010<15		LINEAR		.0001683	DIABETES	MC	service_ca...

Forecasting with Time Series – Exponential Smoothing Node

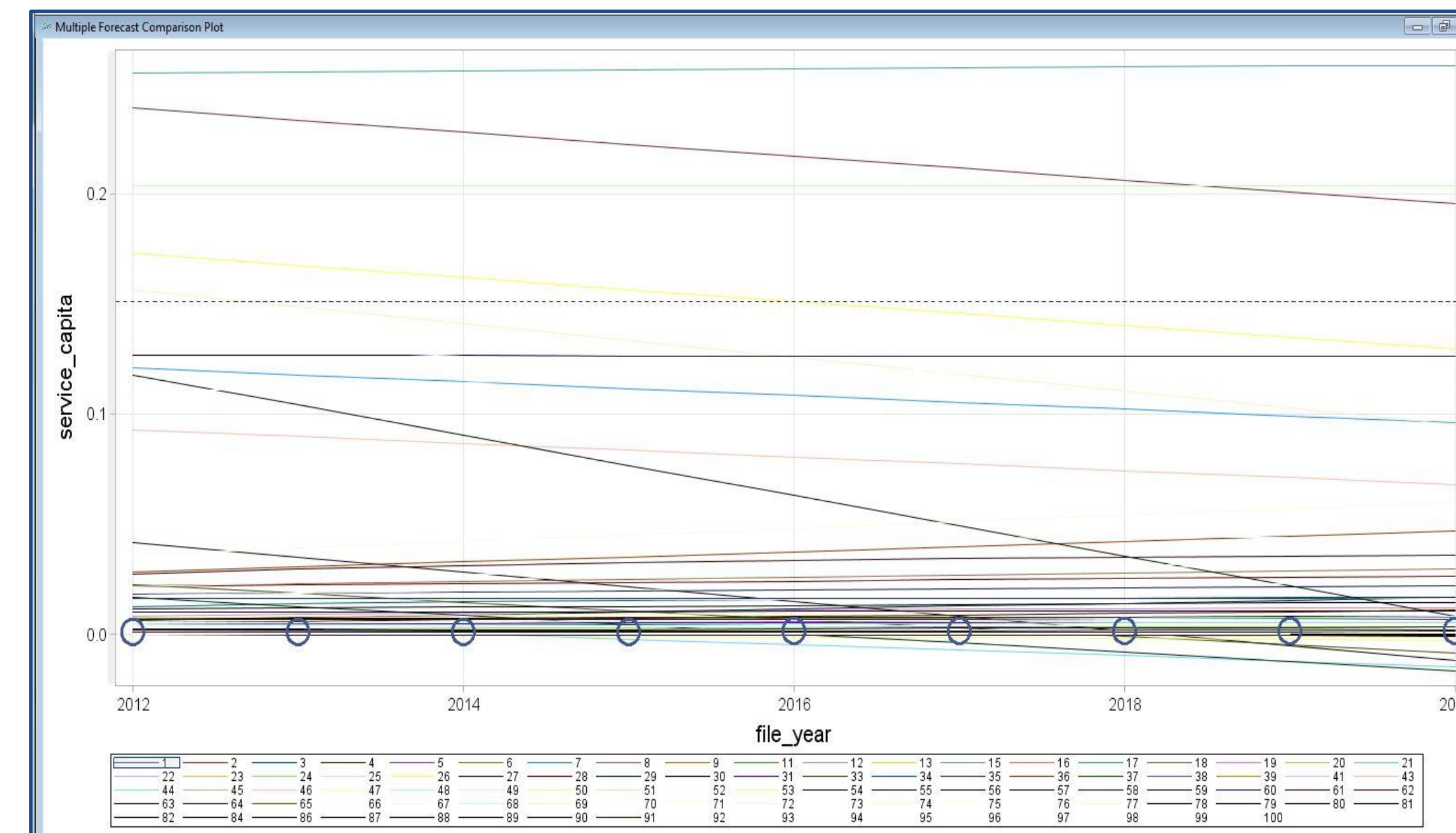
- Cross sectional variables: disease group, age, county, claim type: medical or pharmaceutical
- Model candidates: double, linear, damped trend
- Model selection statistic: mean square error

- Target: services per capita
- Interval: Year
- Accumulation: Total
- Smooth Outliers with predicted value
- Forecasting Method: Best



Property	Value
Train	
Variables	
Specify an Interval	Year
Accumulation	Total
Seasonality	Default
Forecasting Method	Best
Forecast Lead	9
Forecast Back	0
Forecast Sum Start	1
Significance Level	0.05
Input Time Series	
Forecast Input Time Series	No
Extended Value	Predicted Value
Best Model Selection	
Selection Criterion	Mean Square Error
Model Candidates	
Outliers in Exported Data	
Smooth outliers	Yes
Outlier Replacement	Predicted Value
Output Options	
Output Data Type	Coordinated Forecast Data
Lead Point for KLD	1

Property	Value
Simple Exponential Smoothing	No
Double Exponential Smoothing	Yes
Linear Exponential Smoothing	Yes
Damped Trend Exponential Smoothing	Yes
Additive Seasonal Model	No
Multiplicative Seasonal Model	No
Additive Winters Method	No
Multiplicative Winters Method	No



Comparison of Time Series Forecasts

- Baseline Services: 2,178,649
- Concurrent Services Reduction: 110,510
- 5.1% reduction in number of services
- Baseline Cost: \$17,279,931
- Concurrent Programs Savings: \$572,111
- 3.3% reduction in cost





Washington, D.C.
March 23–26, 2014