

Alien Nation: Text Analysis of UFO Sightings in the U.S. Using SAS® Enterprise Miner™ 13.1

Pradeep Reddy Kalakota, FHLBank; Zabiulla Mohammed, Oklahoma State University;
Naresh Abburi, Comcast Corporation;
Dr. Goutam Chakraborty, Oklahoma State University

ABSTRACT

Are we alone in this universe? This is a question that undoubtedly passes through every mind several times during a lifetime. We often hear a lot of stories about close encounters, Unidentified Flying Objects(UFO) sightings and other mysterious things, but we lack documented evidence for analysis of this topic.

The objective of this paper is to analyze one database which has a collection of documented reports of UFO sightings to uncover any fascinating story related to the data. Using SAS® Enterprise Miner™ 13.1, the powerful capabilities of text analytics and topic mining are leveraged to summarize the associations between various reported sightings. We have also used PROC GEOCODE to convert addresses of sightings to the locations on the map and then used PROC GMAP procedure to produce block map to represent the frequency of the sightings in various locations.

Our preliminary analysis of the data associated with sightings, it was found that the most popular words associated with UFOs tell us about their shapes, formations, movements and colors. The Text Profiler node available in SAS® Enterprise Miner™ 13.1 is leveraged to build model and cluster the data into different levels of segment variable. We also explain how the opinions of the UFO sightings changes over a period of time using Text Profiling. Further, this Analysis is used to find interesting terms or topics and model using Text Profile node to describe the UFO Sightings.

INTRODUCTION

Unidentified Flying Objects (UFOs) have been of interest to the public for a long time. Text analysis feature provided by SAS® Enterprise Miner™ 13.1 is a powerful tool to analyze and help interpret textual data about UFO sightings. This paper illustrates various steps involved in the process of analysis and the modeling of textual data for the year 2010 for UFO sightings. The text mining process followed in this paper is the one discussed by Chakraborty, Pagolu and Garla (2014)¹. The scope of this paper is limited to the textual analysis of data, validating the reported information with PROC GEOCODE and producing maps of the sightings across the United States using PROC GMAP. The GEOCODE procedure converts address data to geographic coordinates (latitude and longitude values). These geographic coordinates can then be used on a map in order to calculate distances or also to perform spatial analysis. The application of textual analysis combined with PROC GEOCODE and PROC GMAP procedures can be useful to explore other real-time problems such as location specific sentiment analysis and problem identification.

GEOCODING

The raw dataset has the reported date along with the duration and description of the sighting. This dataset contains all the comments reported from various locations of sightings that were gathered by infochimps website. For analysis we have taken these comments for the year 2010 in the United States. Geocoding is the process of adding geographical coordinates to the addresses, validating the location and deleting duplicate comments. The reported location is validated, converted to geographical coordinates, and the duplicate records were removed. PROC GEOCODE needs two datasets: one is the input dataset and the other one is the lookup dataset. With the new release, SAS help has most of the reference map datasets. From those datasets, the CITY dataset is used as reference in our paper. Frequency of sightings for each city is extracted and used to produce block maps shown in Figure 1.

The codes used for initial data cleaning as well as creating maps are reported in the appendix¹.

GMAP

GMAP procedure is used to produce maps, summarize data that vary by physical area, show trends and variations of data between geographic areas, highlight regional differences or extremes.

Choropleth and block charts on the map were produced with the number of sightings per state. In addition, we pulled in temperature and population data for each state and created maps with this information to understand the effect of external factors on reported number of sightings. California has the highest number of UFO sightings reported. Code for PROC GMAP can be found in appendix².

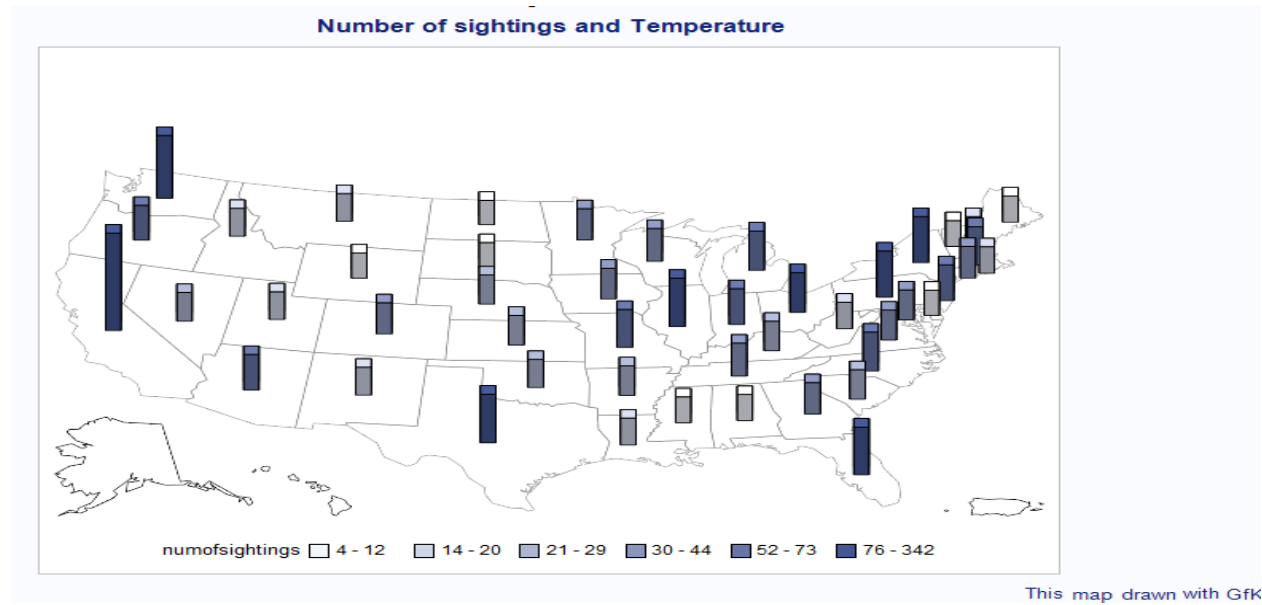


Figure 1. Block Map showing the UFO sightings by each state

Below, block map shows the distribution of the number of sightings per million population generated using PROC GMAP

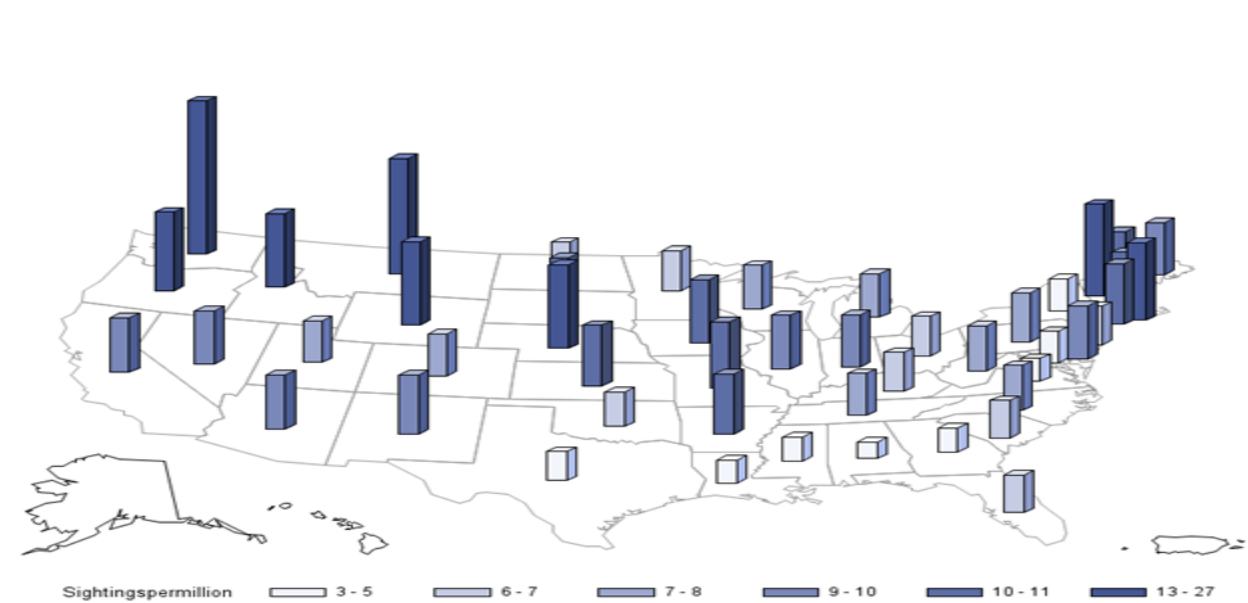


Figure 2. Block Map showing the UFO sightings for 1 million population by each state

The number of sightings per million population per state is calculated by including the census data for the population in the year 2010. The highest number of sightings were reported in the state of Washington.

We have considered temperature effect if it has any impact in the number of sightings. The maximum and minimum temperatures are taken from the National Climate Data Center for the year 2010 recorded in each state. These data points were added to the block map to see the temperature variations across different states during the sightings.

Maximum temperatures across the U.S at the time of the sightings are shown in the map below.

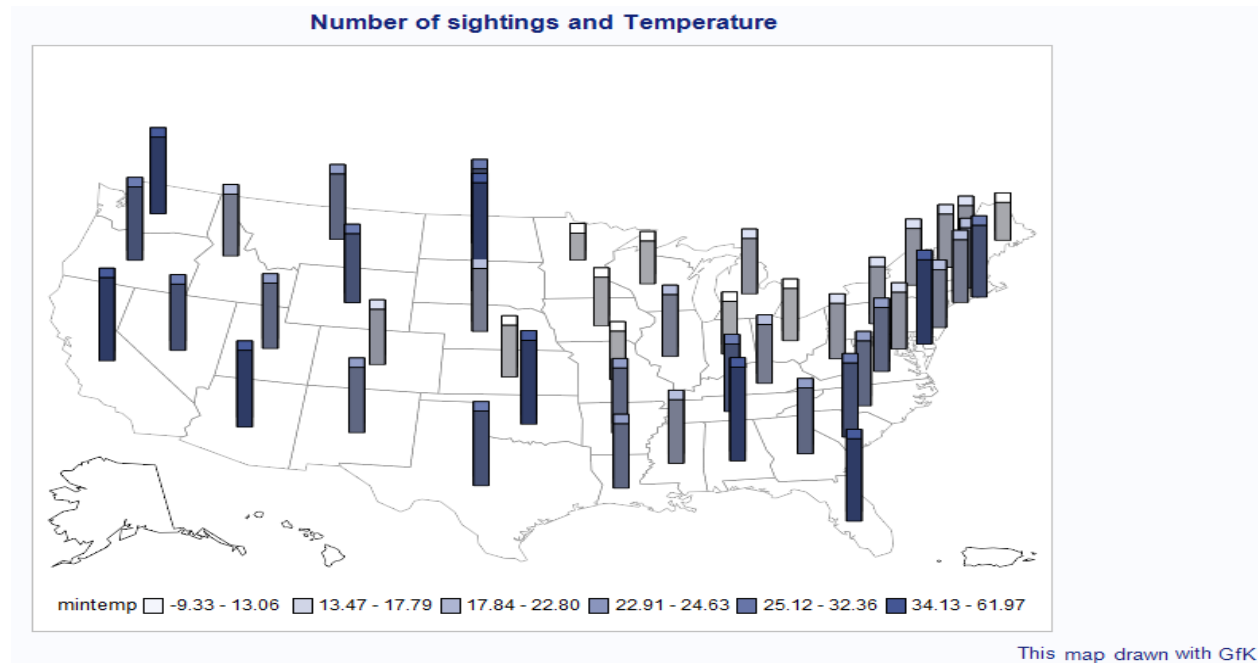


Figure 3. Block Map showing the Maximum temperatures by each state

Minimum temperatures across the U.S at the time of sightings are shown in the map below

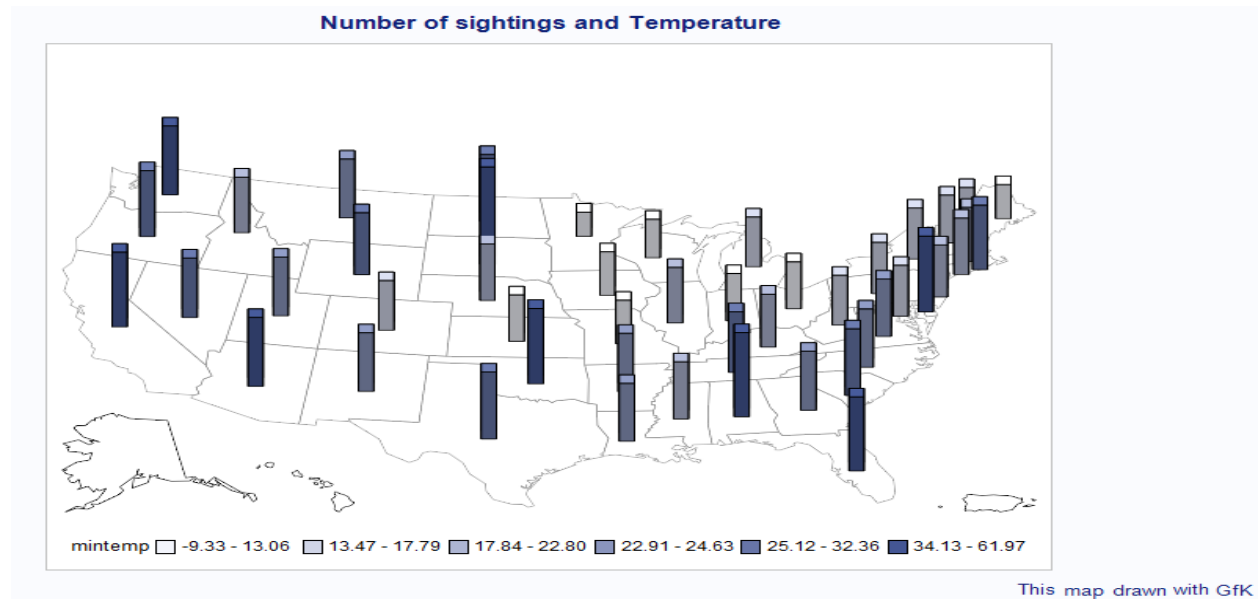


Figure 4. Block Map showing the UFO sightings for 1 million population by each state

From the maps, we can infer that the number of sightings are fewer in places with low temperatures and the maximum number of sightings happens in the state of California that has the temperatures in the range of 25°- 60°F.

DATASET

UFO Sightings 2010 Dataset:

SAS Variable Format	Data Type	Data Source
UFOSightingDescription	CHARACTER	infochimps
UFOSightingLocation	CHARACTER	infochimps
UFOSightingDate	DATE	infochimps
UFOReportingDate	DATE	infochimps
Comments	CHARACTER	infochimps

Table 1. UFO sightings dataset: Data Types for SAS Variable Formats

Climate dataset:

SAS Variable Format	Data Type	Data Source
DATE	DATE	National Climatic data Center
Location	CHARACTER	National Climatic data Center
Max Air Temperature	NUMERIC	National Climatic data Center
Min Air Temperature	NUMERIC	National Climatic data Center
State	CHARACTER	National Climatic data Center

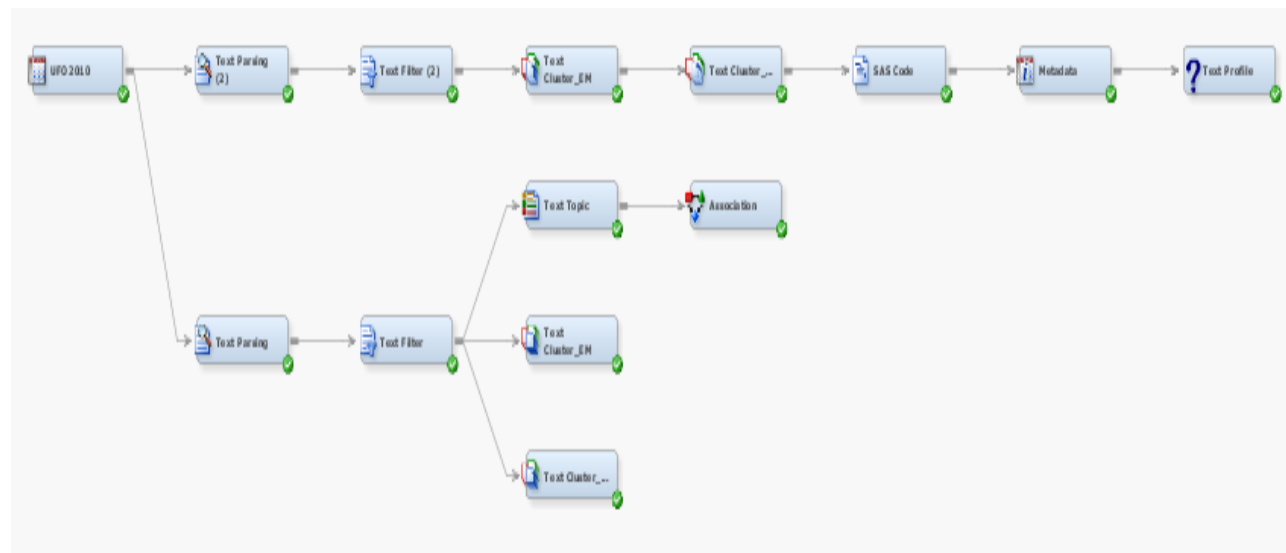
Table 2. Climate Dataset: Data Types for SAS Variable Formats

Population Dataset:

SAS Variable Format	Data Type	Data Source
State	CHARACTER	Census.gov
Population	NUMERIC	Census.gov
Year	DATE	Census.gov

Table 3. Population dataset: Data Types for SAS Variable Formats

PROCESS FLOW



Display 1. Text mining process flow in SAS Enterprise Miner

TEXT PARSING:

The Text Parsing node enables you to parse a document collection in order to quantify information about the terms. The Text Parsing node is used to parse the text data based using different parts of speech and noun groups. A few words are dropped by reviewing the words and their importance. Following table shows the list of terms that were discarded/kept based on their importance.

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ be	...	Alpha	19003	2574N		+	296	1
+ see	...	Alpha	6409	2135N		+	178	2
+ light	...	Alpha	7853	2012Y		+	45	3
+ sky	...	Alpha	3778	1760Y		+	41	4
+ move	...	Alpha	3395	1534Y		+	206	5
+ look	...	Alpha	3056	1521Y		+	100	6
then	...	Alpha	3054	1475N			118	7
+ have	...	Alpha	3205	1460N		+	169	8
+ no	...	Alpha	2539	1458N		+	192	9
not	...	Alpha	2921	1398N			68	10
+ object	...	Alpha	4247	1379Y		+	252	11

Output 1: List of terms parsed

TEXT FILTERING:

Text Filter node is used to reduce further, the total number of parsed terms or documents that are analyzed. Therefore, we eliminated extraneous information so that only the most valuable and relevant information is considered. User defined synonym list is created using interactive filter to give a definitive name which can identify a set of words to generalize the terms. A few words are dropped by reviewing the words and their importance. Below table shows the exported synonym list that was created to use.

	term	termrole	parent
1	nigh		night
2	tiny airport	NOUN_GROUP	tiny
3	tiny amount	NOUN_GROUP	tiny
4	tiny bit	NOUN_GROUP	tiny
5	tiny blink light	NOUN_GROUP	tiny
6	tiny blue light	NOUN_GROUP	tiny
7	tiny dot	NOUN_GROUP	tiny
8	tiny dot-don	NOUN_GROUP	tiny
9	tiny fraction	NOUN_GROUP	tiny
10	tiny glow dot	NOUN_GROUP	tiny
11	tiny light	NOUN_GROUP	tiny
12	tiny light blue	NOUN_GROUP	tiny
13	tiny lightwitnessed	NOUN_GROUP	tiny
14	tiny little line	NOUN_GROUP	tiny
15	tiny little red	NOUN_GROUP	tiny
16	tiny municipal airport day	NOUN_GROUP	tiny

Display 2. Synonyms list in text parsing node

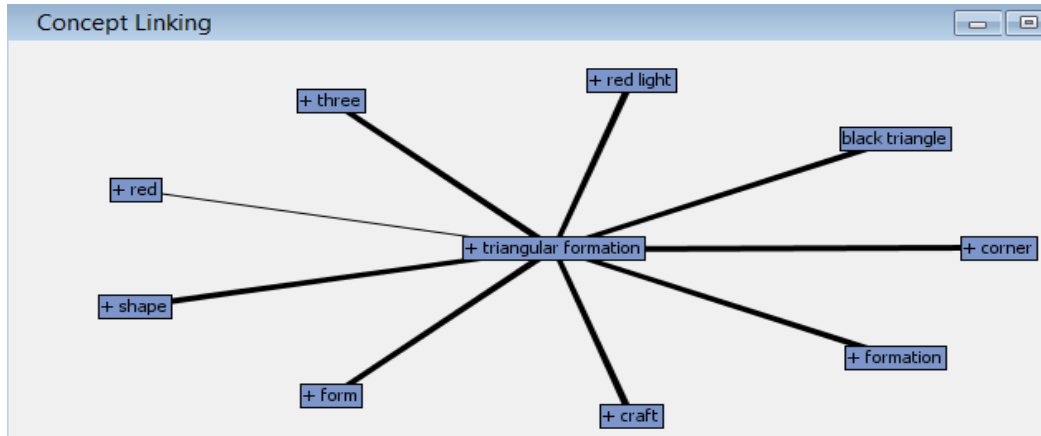
The Text Filter node assigns weight to the words based on their relative frequencies. The default setting was used to filter all the words which have a low weight. Check spelling was set to 'Yes'. SAS default weighting was used to calculate weighted frequency for terms. A term by document matrix was created with terms and their respective weight

Term	Role	Attribute	Status	Weight ▼	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank
+ curtain	...	Alpha	Keep	0.811	10	10	6	6	2706
+ probe	...	Alpha	Keep	0.806	11	14	5	6	2706
+ teenager	...	Alpha	Keep	0.806	12	12	6	6	2706
+ luminescence	...	Alpha	Keep	0.797	7	10	4	6	2706
oclock	...	Alpha	Keep	0.797	10	10	6	6	2706

Output 2: Term by Document Matrix

CONCEPT LINKS:

Undeniably, the coolest feature of Text Mining in SAS Enterprise Miner is the Concept Link Diagram. Concept Link diagrams are visual representations of how terms are related to one another. When we generated concept link diagrams on the UFO sighting reporting we generated many interesting concept links such as the one below:



Output 3: Concept Link Diagram

UFO theories generally categorize the appearance of the objects as many geometric formations and the most popular formation is the “Triangular Formation.” From the comments, the concept link generated above shows various terms associated with the term “Triangular Formation.” The thickness of the link between the terms displays the strength of the relationship. The following is inferred from the concept link diagram:

1. The strongest association (the thickest link) is with the word “corner.”
2. The association with the term “black triangle” explains to us that the word “black triangle” appears most of the time in the sightings that talk about “triangular formulations.”
3. Another significant association is found with the term “red light.” This is due to the fact that the sighting comments have reported that the triangle shaped UFOs they saw often appeared with red lights. We can also observe another weak association with the color “red.”

TEXT TOPICS:

The Text Topic node enables you to create topics of interest from a list of terms. Text Topic node extracts topics from text documents. A document can have multiple topics. Parameters of the text topic node can be manually set based on the size of the data and the frequency of unique terms. For this project the numbers of terms in a topic are limited to 15. Below term topic matrix shows us the distribution of various terms together as topics.

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.632	0.082	contact information,anonymous,+elect,contact,information	82	434
Multiple	2	0.793	0.083	+star,+satellite,+bright,+light,+white	290	440
Multiple	3	0.772	0.084	+house,+window,+hear,+run,+look	375	400
Multiple	4	0.733	0.093	+road,+car,+drive,+light,+pull	329	354
Multiple	5	0.712	0.082	+object,+appear,aircraft,+observe,approximately	384	424
Multiple	6	0.614	0.072	orange,+orange,glowing,+ball,+orange light	287	432
Multiple	7	0.518	0.071	rsquo;dquo,didn,+thing	291	316
Multiple	8	0.489	0.068	+photo,+picture,+camera,+video,+ufo	297	358
Multiple	9	0.534	0.070	aircraft,+fly,+plane,+jet,+hear	291	418
Multiple	10	0.530	0.068	+firework,july,4th,+fireball,orange	285	333
Multiple	11	0.559	0.069	+object,+shape,+plane,+cloud,+shape	323	429
Multiple	12	0.529	0.067	+green,+flash,+light,+red,flashing	253	412
Multiple	13	0.522	0.065	+formation,+triangle,three,+light,+light	281	426
Multiple	14	0.502	0.064	west,east,north,south,+degree	272	427
Multiple	15	0.430	0.063	+date,+report,+sighting,+indicate,nuforc	263	387
Multiple	16	0.476	0.065	+apos,+satellite,quot,+line,+binocular	381	394
Multiple	17	0.429	0.064	+apos,+ufo,+know,+thing,didn	308	426
Multiple	18	0.457	0.061	+red,+red light,+white,+white light,+red	266	391
Multiple	19	0.416	0.061	+craft,+shape,+large,+shape,triangular	235	387
Multiple	20	0.404	0.059	+fireball,+meteor,+fall,+star,green	294	403
Multiple	21	0.318	0.057	+object,+fireball,+apos,+fly	192	369
Multiple	22	0.332	0.056	+cloud,+flash,+cover,+orb,+white	322	405
Multiple	23	0.351	0.056	+tree,+tree,+bright,+light	299	438
Multiple	24	0.328	0.056	+tree,+tree,+straight,+straight line,+tree line	263	382
Multiple	25	0.309	0.054	+line,+tree,+straight,+straight line,+tree line	263	382

Output 4: Text topics

The above table is shown in the below matrix as distribution.



Output 5: Topic matrix

Based on the document cutoff we can say that the topics with high frequency are about the initial thoughts on the shape and the color of the aircrafts when people spotted UFOs. From the above topic ID 9, we can find that the UFO is similar to an aircraft with a peculiar visibility as they disappear in seconds and people either think they saw a star or satellite with a glowing object effect. Text topic 4 says that most of the people noticed UFO's while they commute to home from their office which implies that they occur mostly during evening or late in the nights. Text topic analysis using enterprise miner is one way of obtaining useful information from the initial analysis of comments.

TEXT CLUSTERING:

Text Clustering assigns each document to a cluster and use Singular Value Decomposition (SVD) to reduce the curse of dimensionality. The Max SVD dimension property was set to 40 and max terms in a cluster are set to 15 with the number of clusters limited to 12.

Two algorithms Expectation-Maximization and Hierarchical clustering methods were used and compared to explore if the clusters make sense. Hierarchical clustering seemed to create more meaningful clusters with this data and therefore used further to model using Text Profile node.

Hierarchical Clustering:

Cluster ID	Descriptive Terms	Frequency	Percentage
11	+bright +star stars +north +night seconds moving +fast +light +east +speed +south +sky +direction white	398	14%
12	+contact information' +contact +fire +information +orange +remain +totally +witness anonymous elects fireworks +note +video +camera witnessed	122	4%
16	+contact information' +contact +information +note +remain +totally +witness anonymous elects +bright +green +light +down +star +car	438	16%
17	+craft +picture +red driving hovering lights triangle shaped +shape +camera +home +ufo +video +three +back	334	12%
18	white turned +note +light anonymous elects +contact information' traveling +remain +south +totally +witness +information +green +contact	55	2%
19	+don +know +craft +report +house flew +ufo outside +thing flying shaped feet +air +look +shape	152	6%
20	+orange fireworks driving +car +south +home +red +line appeared +slowly +area +approximately +north disappeared minutes	265	10%
21	+orange +house later +line lights +red +back +large +night +slowly moved +three +clear few disappeared	105	4%
22	+craft feet flying +low driving +aircraft flew +air shaped +altitude traveling +large +sound +speed +car	237	9%
23	+orange +approximately appeared west +south +straight +speed +altitude +north +line +east +sound traveling +direction seconds	234	9%
24	few left appeared moved +move watched +right +slowly shaped noticed +down +three started minutes seconds	157	6%
27	+altitude +aircraft +clear +approximately +high +speed +object +sound traveling +air west shaped +craft feet +east	248	9%

Output 6: Hierarchical Clusters

From the clustering output, we have named the four most important clusters which categorize about 52% of the comments as follows:

Cluster ID 16 – The Travelers

This cluster has the highest number of documents in it. This cluster tells us about the people who witnessed anonymous objects flying with a bright green light. Most of them witnessed the UFOs as lights when they were travelling in a car.

Cluster ID 11 – The Navigators

This cluster has the second highest number of documents. This cluster has the comments in which bright stars appeared during the night. Also, they have reported about the direction of their appearance such as bright stars appearing from the north/south and moving towards the east at a fast pace.

Cluster ID 17 – The Geometrists

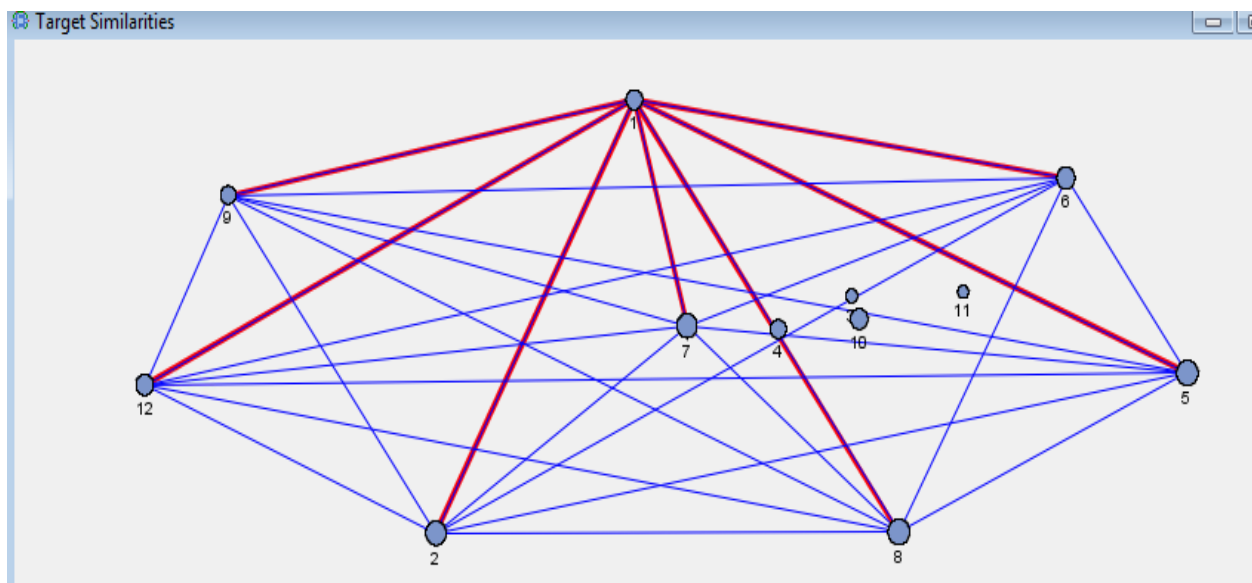
This cluster has the third highest number of documents. This cluster includes those comments which spoke about the specific shapes of the UFO such as “a triangle shaped aircraft.” They also spoke about the number of lights they observed, such as “three lights observed from their backyards.”

Cluster ID 20 – The Observers

This cluster has the fourth highest number of documents. This cluster includes comments about UFOs which appeared as a line and have described them as “fireworks” when they were driving or when they were at home. Also, they reported about the direction in which the UFO appeared and disappeared.

TEXT PROFILING:

Text Profiler is a new node only available from SAS Enterprise Miner 13.1. The Text Profile node enables to profile a target segment (cluster level) variable using terms found in the documents. The approach uses a hierarchical Bayesian model to predict which terms are the most likely terms to describe the level. For each level of a target segment variable, the node outputs a list of terms from the collection that characterize or describe that level. This node enables us to model the text data into different cluster levels. The 12 clusters created using hierarchical clustering algorithm and the target variable is set as the segment variable with 12 levels. The reported terms for level n are compared to those at level n-1. The following target similarities give us about the relationship between various clusters created using Hierarchical cluster node.

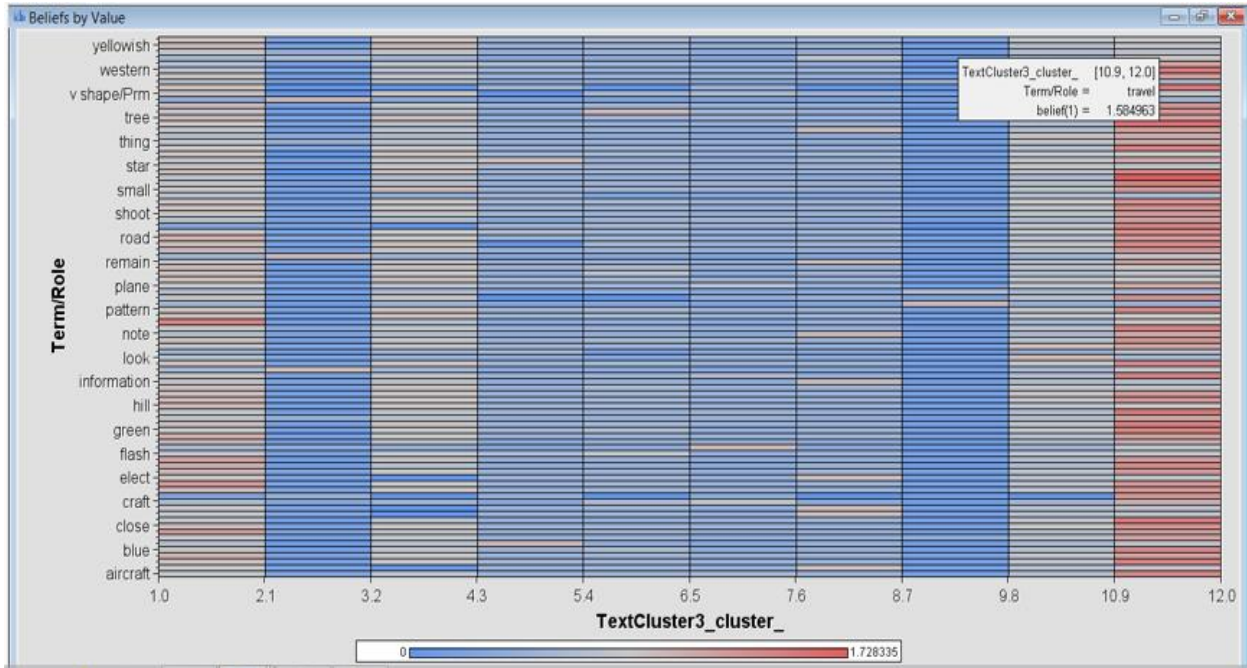


Output 7: Target Segments

Target Similarities diagram explains if there is any relationship between clusters. Cluster 11 is not connected to any of the other clusters because it does not share any similarities. Cluster 11 consists of the

terms “speed, high, travel, cloud, vapor, sound, sun, jet”. If we clearly observe the terms in cluster 11, they are not similar to the terms in other clusters. This is why cluster 11 is isolated.

This is a diagrammatic representation of 12 different segments and their weights and the terms associated with it. The Beliefs by Value graph shows belief values for term and role pairs for various target values. For the term “travel”, the belief value is given as 1.5849 which profiles this term in cluster 11.



Output 8: Target Similarity Diagram

More about these terms and the profiling results are shown in the below table. From, the level of cluster and their profiled terms, it is clear that a cluster is profiled to make a meaning out of terms.

Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Freq	Corpus	TextCluster3_cluster_	Var/Var Level ID	Terms
Corpus		red lightNg	aircraft	highway	vehicle	east	sky	green	light	2745	1			0red lightNg, aircraft, highwa...
TextCluster3... 1		drive	car	road	right	home	hill	down	turn	215		1		1drive, car, road, right, home...
TextCluster3... 2		orange	firework	glow	fireball	ball	sky	fire	red	316		2		2orange, firework, glow, fireb...
TextCluster3... 3		report	ufo	know	witness/Prm	deg	hear	wrong	thing	66		3		3report, ufo, know, witness/P...
TextCluster3... 4		tree	yellowish	small	pulsate	orange light...	stationary	river	spherical	187		4		4tree, yellowish, small, pulsa...
TextCluster3... 5		bright	star/Til	satellite	flash	white object...	star	move	shoot	357		5		5bright, star/Til, satellite, flas...
TextCluster3... 6		triangular for...	flash	red	blue	craft	red lightNg	green	hover	252		6		6triangular formationNg, flas...
TextCluster3... 7		fly	craft	plane	jet	v shape/Prm	aircraft	shape	noise	319		7		7fly, craft, plane, jet, v shape/...
TextCluster3... 8		note	contact	totally	contact infor...	anonymous	information	elect	remain	356		8		8note, contact, totally, contact...
TextCluster3... 9		photo	picture	video	camera	photograph	shoot	zoom	show	167		9		9photo, picture, video, camer...
TextCluster3... 10		look	watch	night	house	time	close	star	sit	232		10		10look, watch, night, house, ti...
TextCluster3... 11		speed	high	travel	cloud	vapor	sound	sun	jet	41		11		11speed, high, travel, cloud, v...
TextCluster3... 12		formation	group	approximately	observe	western	shadow	pattern	location	237		12		12formation, group, approxim...

Output 9: Cluster Model

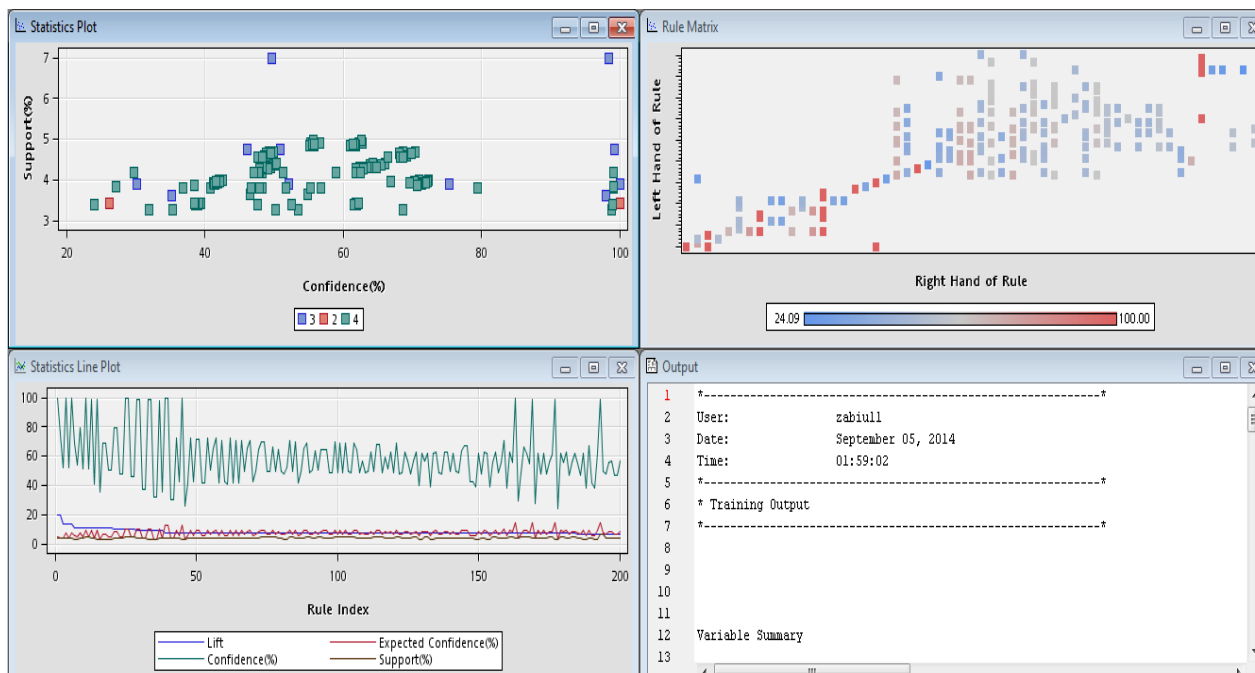
This above table shows various terms that best predicts each cluster. These are useful to get the following interpretation of the clusters.

- 1) Profiles related to people driving home and comments about their directions when they encounter UFO sightings will have high probability to fall into cluster 1.
- 2) Profiles related to pattern, formations fall under cluster 12 as people commented about the pattern of UFO's and various shapes of the formations.

Using Target similarities we can profile new data into various clusters which have meaningful information from the trained data.

ASSOCIATION ANALYSIS:

Association discovery is the identification of items that occur together in a given event or record. Association discovery rules are based on frequency counts of the number of times items occur alone and in combination. Association rules define some affinity between two or more items. Text topic node is used as an input For example, according to one of the association rules, documents with the term "bizarre" are associated with the document that about police investigation with a confidence of 78 percent. The term "fire ball" with "orange" has a confidence of 86 percent with a lift of 7.08. Statistics of these rules in the rule matrix are shown in the statistics plots below.



Output 10: Association Analysis

CONCLUSION

This text mining project was intended to find interesting concepts associated with UFO sightings and profile them using the powerful features available in SAS Enterprise Miner™ 13.1. External data exploratory analysis combined with text analytics provided a sound understanding of text data. PROC GEOCODE helped in validating the location of the comments posted. Powerful map utility PROC GMAP provided visual analysis on the map with projected blocks with the frequency of sightings. Temperature has some effect on the number of sightings reported and the population has a major impact in reporting the UFO sightings.

Our text analysis reveals that UFOs are similar to an aircraft with a peculiar visibility and they disappear within seconds after their appearance. UFOs are generally triangular in shape with a red light and, in some cases, it is visible as a black triangle or other crafts. Text clustering revealed some interesting facts. Most of the UFO reporters have observed the UFOs either when they were driving in their cars or when they are at their homes. Also, reporters have used the term "fireworks" to describe UFOs. UFOs were reported as appearing in the north or south direction and disappearing in the north. Based on our analysis, we found that most UFO encounters are surprising, fun, and unforgettable experiences.

REFERENCES

Goutam Chakraborty, Murali Pagolu, Satish Garla. 2014. "Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®" SAS. Available at <http://support.sas.com/publishing/pubcat/chaps/65646.pdf>

Stephen McDaniel, Chris Hemedinger. March 2010. "SAS® For Dummies®, Second Edition". Available at http://www.sas.com/store/prodBK_62824_en.html

Darrell Massengill and Ed Odom. Paper 332-2010 "PROC GEOCODE: Now with Street-Level Geocoding". Proceedings of SAS Global Forum 2010, SAS Institute Inc., Cary, NC

SAS PROC GEOCODE Documentation. Available at <http://support.sas.com/documentation/cdl/en/graphref/63022/HTML/default/viewer.htm#a003121448.htm>

SAS PROC GMAP Documentation. Available at <http://support.sas.com/documentation/cdl/en/graphref/63022/HTML/default/viewer.htm#gmap-proc-statement.htm>

ACKNOWLEDGMENTS

We thank Dr. Goutam Chakraborty, Founder of SAS and OSU Data mining certificate program, Oklahoma State University, for his support, guidance and encouragement throughout our research.

AUTHORS

Pradeep Reddy Kalakota was a Graduate Student at Oklahoma State University and completed OSU and SAS data mining certification program. He is currently working as Business Intelligence Developer II for Divisional Risk team at FHL Bank Des Moines. He is a Base SAS® 9 certified Programmer, SAS® Certified Statistical Business Analyst, SAS® certified Predictive Modeler using Enterprise Miner 7. He has 5 years of experience in ETL-Data Warehousing at Deloitte Consulting.

Zabiulla Mohammed is a Masters' student in Management Information Systems at Spears School of Business, Oklahoma State University. He holds SAS Statistical Business Analyst, Predictive Modeler, Base and Advanced Programmer for SAS 9 Credentials. He has 5 years of experience working with two Fortune 100 companies. He successfully completed SAS and OSU Data Mining Certificate program in December 2014. He has an undergraduate degree in Computer Science and Engineering and has presented various papers at conferences including JMP Discovery Summit 2014 and SAS Analytics Conference 2014

Naresh Abburi is a MIS graduate from Spears School of Business at Oklahoma State University. He is a SAS certified Statistical Business Analyst, SAS certified Base SAS Programmer and SAS certified Advanced SAS Programmer. He received SAS & OSU Data mining Certificate in May 2014. Currently, he is working as a Senior Analyst - Customer Analytics at Comcast, Philadelphia. He has worked for IBM & Headstrong Capital Markets for 4.5 years.

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU marketing analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Pradeep Reddy Kalakota
Federal Home Loan Bank of Des Moines
pradeep.kalakota@okstate.edu
<https://www.linkedin.com/in/pradeepreddykalakota>

Zabiulla Mohammed
Oklahoma State University
zabiulla.mohammed@okstate.edu
<https://www.linkedin.com/in/zabiullamohammed>

Naresh Abburi
Oklahoma State University
naresh.abburi@okstate.edu
<https://www.linkedin.com/in/nareshabburi>

Dr. Goutam Chakraborty
Oklahoma State University
Stillwater, OK, 74078
goutam.chakraborty@okstate.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX:

USING PROC GEOCODE:

```
PROC SQL;          /* JOIN WITH US MAP DATA TO VALIDATE LOCATIONS*/
  CREATE TABLE WORK.QUERY_FOR_UFO2010S_SAS7BDAT_0000 AS
  SELECT T1.CITY
    FROM WORK.QUERY_FOR_UFO2010S_SAS7BDAT T1
      INNER JOIN MAPSSAS.USCITY T2 ON (T1.CITY = T2.CITY)
  WHERE T1.CITY NOT = '';
RUN;
QUIT;

PROC GEOCODE
  METHOD=CITY          /* SPECIFY GEOCODING METHOD          */
  DATA=MAPDATA.UFO2010 /* INPUT DATA SET OF CITIES          */
  OUT=GEOCODED_DATA; /* OUTPUT DATA SET WITH LOCATIONS */
RUN;
QUIT;

PROC SQL;          /* REMOVE THE DUPLICATE AND UNMATCHED LOCATIONS*/
  CREATE TABLE WORK.FILTER_FOR_ABC AS
  SELECT T1.LAT,
    T1.LONG,
    T1.M_OBS,
    T1._MATCHED_,
    T1.UFOSIGHTINGLOCATION,
```

```

        T1.CITY,
        T1.STATE
    FROM WORK.ABC T1
    WHERE T1.LAT NOT = .;
QUIT;

```

USING GMAP CODE:

```

PROC SQL;                                /* JOIN WITH CLIMATE DATA          */
    CREATE TABLE WORK.QUERY_FOR_MAINMAIN AS
    SELECT T1.'UFOSIGHTINGLOCATION 2'N AS LOCATION,
           T1.UFOSIGHTINGDATE,
           SUBSTR(T1.UFOSIGHTINGDATE,3,3) AS MON,
           T2.STATE,
           T2.'AVG DAILY MAX AIR TEMPERATURE (F'N AS TEMPMAX,
           T2.'AVG DAILY MIN AIR TEMPERATURE (F'N AS TEMPMIN
    FROM WORK.MAINMAIN T1
         LEFT JOIN WORK.MAINEXCEL T2 ON (T1.UFOSIGHTINGDATE = T2.'MONTH
DAY, YEAR'N) AND (T1.'UFOSIGHTINGLOCATION 2'N
                = T2.STATECODE)
    WHERE STATE IS NOT NULL;

CREATE TABLE WORK.FINAL AS              /* CALCULATE AVG MAX AND MIN TEMP    */
SELECT  STATE,
        MON,
        COUNT(*) AS NUM,
        AVG(TEMPMAX) AS MAXTEMP,
        AVG(TEMPMIN) AS MINTEMP
FROM WORK.QUERY_FOR_MAINMAIN
GROUP BY STATE, MON;

CREATE TABLE WORK.FINALSTATE AS        /* GET TEMPERATURE BY STATE          */
SELECT  STATE,
        MIN(MAXTEMP) AS TEMPMAXIMUM,
        MAX(MINTEMP) AS TEMPMINIMUM
FROM WORK.FINAL
GROUP BY STATE;

QUIT;

PROC SQL;
    CREATE TABLE WORK.FILTER_FOR_FINALFINAL1_SAS7BDAT AS
    SELECT T1.STATECOD,
           T1.MON,
           T1.NUM,
           T1.MAXTEMP,
           T1.MINTEMP,
           T1.STATE,
    FROM EC100044.FINALFINAL T1
    ORDER BY T1.STATECODE;

QUIT;

PROC SQL;                                /* GET STATE CODE NUMBER FROM MAPS    */
    CREATE TABLE WORK.FILTER_FOR_US AS
    SELECT DISTINCT T1.STATE,
           T1.STATECODE
    FROM MAPSGFK.US T1;

```

```

QUIT;
PROC SQL;                                /* PREPARE DATA*/
  CREATE TABLE WORK.FILTER_FOR_FINALFINAL_SAS7BDAT AS
  SELECT T1.STATECOD,
         T1.MON,
         T1.NUM,
         T1.MAXTEMP,
         T1.MINTEMP,
         T1.STATE AS STATECODET1,
         T2.STATECODE AS STATECODET2,
         T2.STATE AS STATE
  FROM EC100044.FINALFINAL T1, WORK.FILTER_FOR_US T2
  WHERE T1.STATE=T2.STATECODE
  ORDER BY T1.STATE;
QUIT;

PROC SQL;                                /* PREPARE FINAL DATASET */
  CREATE TABLE WORK.FILTER_FOR_FINALFINAL_SAS7BDAT AS
  SELECT SUM(T1.NUM) AS NUMOFSIGHTINGS,
         MAX(T1.MAXTEMP) AS MAXTEMP,
         MIN(T1.MINTEMP) AS MINTEMP,
         T1.STATE
  FROM WORK.FILTER_FOR_FINALFINAL_SAS7B_0000 T1
  GROUP BY T1.STATE
  ORDER BY T1.STATE;
QUIT;

GOPTIONS RESET=ALL BORDER;              /* OPTIONS FOF BUILDING BLOCKMAP */
TITLE1 "NUMBER OF SIGHTINGS AND TEMPERATURE";
LEGEND1 LABEL=(POSITION=TOP) SHAPE=BAR(.3IN,.1IN);
FOOTNOTE1 J=R "THIS MAP DRAWN WITH GFK MAP DATA";
PATTERN1 VALUE=MSOLID COLOR=TAN;

/* PROC GMAP TO CREATE BLOCK MAPS ON MAP*/
PROC GMAP DATA=WORK.FILTER_FOR_FINALFINAL_SAS7BDAT MAP=MAPS.US ALL;
ID STATE;
CHORO NUMOFSIGHTINGS MAXTEMP MINTEMP/ DISCRETE LEGEND=LEGEND1
COUTLINE=GRAY98 ANNOTATE=LABELOUT;
BLOCK NUMOFSIGHTINGS MAXTEMP MINTEMP / BLOCKSIZE=2 RELZERO;
RUN;

```