SI601 Final Project Report

Name: Kritika Moolchand Versha

Motivation

As defined in Wikipedia, light pollution competes with starlight in the night sky for urban residents, interferes with astronomical observatories, and, like any other form of pollution, disrupts ecosystems and has adverse health effects. It is most severe in highly industrialized, densely populated areas of North America, Europe, and Japan and in major cities in the Middle East and North Africa like Tehran and Cairo, but even relatively small amounts of light can be noticed and create problems. Since the early 1980s, global dark-sky movement has emerged, with concerned people campaigning to reduce the amount of light pollution.
Globe at Night has been collecting data about the night sky since 2006.By using this data; my goal is to find how light pollution has affected the visibility of constellation with naked eyes in subsequent years. Higher the limiting magnitude of measurement, lower is the visibility.

Data Sources

Data Source 1: **Description:**
The dataset available at Globe At Night Campaigns contains description of the visibility of constellations capturing details about the location from where the data collected along the measured visibility in terms of limiting magnitude and comments on the sky whether it was clear or not.
Location of data source: http://www.globeatnight.org/maps.php
Format: .csv files
Important variables:

- Observation Date -> Column D ->ObsDate
- Limiting Magnitude -> Column F ->Limiting Mag
- Latin Name of the Constellation -> Column J ->Constellation
- Country ->Column M

Time Period of the data collected: Year 2010-2013 with each year, observations taken during the first half of the year. Each constellation has a visibility month which improves the accuracy of observation which was missing from the dataset from year 2006-2009, hence was not used for processing
Total Number of records:
47439

Data Source 2 Description: This page provides a detailed description about the 88 identified constellations visible from earth with naked eyes
Location of data source: http://www.nightskyatlas.com/constellations.jsp
Format: JSP page
Important variables:

- Latin Name of Constellation
- Number of principal stars
- Best visibility month

Time Period of the data collected: Not Applicable
Total number of records: 88

Data Source 3 Description: The data source was collected by Paul Rodmell which describes the calendar months with the visibility of constellations. It is available at the Royal Astronomical Society of New Zealand

Location of data source: http://www.rasnz.org.nz/Stars/Constellations.shtml#months

Format: HTML page

Important variables:

- Months
- Visible Constellations

Time Period of the data collected: 12 months

Total number of records: 12

I have used the names data source 1, data source 2 and data source 3 to provide descriptions of the data for the rest of the report

Data Manipulation

Data source was converted in the following format:

Data Source 1: The .csv files available was loaded into SQL tables using python and sqlite

Data Source 2: The jsp page was converted into a .csv file

Data Source 3: the html page was converted into a .csv file

Missing/incomplete data:

The dataset from Globe at Night from 2006-2013 was not in the same format. Some of the key information like the name of the observed constellation was missing from the dataset from year 2006-2009.This variable plays an important role in measuring the accuracy of visible magnitude.

For example, if you try to observe a constellation during a time period which is not the best for it to be visible; the measurement of limiting magnitude gets affected. Steps which I took to handle missing data are as follows:

- I did not load the observations from the year 2006-2009 due to absence of observed constellation in the file
- The limiting magnitude is measured on the scale of 0 to 7.There were records which had no detail of the magnitude or there were records which had an out of range number in this variable. Also, there were records which had missing name of constellation in it.
- Also , there were observations made with constellation during the time period which was not their best observation period.
  I filtered such data using conditions in the where clause query while processing the data

Processing:

Step 1: *Creation of table in SQL*

I created three SQL tables in the database using sqlite

- ‘light’ - Observations from data source 1
- ‘constellation’ using data source 2
- ‘visibility_months’ using data source 3

Step 2:Manipulation of table observation details in table 'light'

- The observation date from table 'light' was used to add a column 'Year' to the same table
- The months in textual format was concatenated with the observation date column in the light table
  For example: if the observation date is 06/21/2013, the string was formatted to 06/21/2013-June

Step 3: Joining of tables

- The tables were combined using 'JOIN'
- Incomplete data / noise was filtered using where clause
  - Records with out of range limiting magnitude was filtered
  - Records with missing name of constellation was filtered
- The constellation name and observation date from table 'light' was used to match with the constellation name and months from the table 'visibility_months' to filter the observation records which could affect the calculation of limiting magnitude due to incorrect mapping of visibility months and constellation stars
- The result was grouped using the month and year of observation with average value of limiting magnitude for each year/each month of the year calculated from available data.

Workflow of the source code:

- Proj2_File.py was used to extract the all the fields from data source 2 into html file *Pro_step1.html* and then adding them into tab separated *Proj2_File.txt* file
- Proj5_File.py was used to extract the all the fields from data source 3 into html file *Pro_step2.html* and then adding them into tab separated *Proj5_File.txt* file
- Proj3_file.py was used to load the dataset from data source 1, *Proj2_File.txt* file and *Proj5_File.txt* file into SQL tables
- Proj1_file.py was used to update and manipulate the 'light' table
- Proj4_File.py was used to combine the datasets and obtain comparison results which were manually copied into the file.

Solving Challenges:

I had to think long as how would I use the observation date from the datasets to match it with the months from visibility_month table and how would I group the result according to the months and year. It was a challenge because of the format of the observed date was different from the month's column of the visibility_month.

I assigned a month name and concatenated it with the observation date field according to the given month number in the field so I could get a common format in substring for matching patterns. Also, to group the data according to the month, I added a column named 'Year' which was the year extracted from the observation date field to make grouping easy

Analysis and Visualization

Key goal in combining data sources:

Given that the limiting magnitude is provided in the dataset, it could have been said that this information would have been suffice to calculate the average limiting magnitude for each month and hence each year. The Globe At Night Campaign uses the app. http://www.globeatnight.org/webapp/ to gather data from the users each year to calculate the limiting magnitude of the sky who are allowed to use Orion, Leo and Crux to capture details.

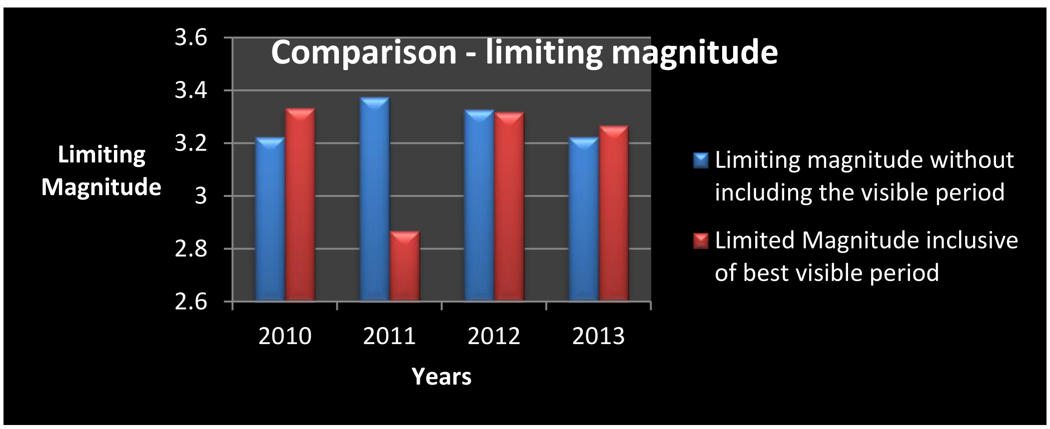
I combined Globe at Night dataset with the dataset from the Royal Astronomical Society of New Zealand which shows a calendar of twelve months, each month showing visibility of constellation.

Limiting magnitude is a measure of sky brightness. A faded constellation would not be able to provide accurate details about limiting magnitude and hence combining both the dataset is worth to give accurate results about limiting magnitude

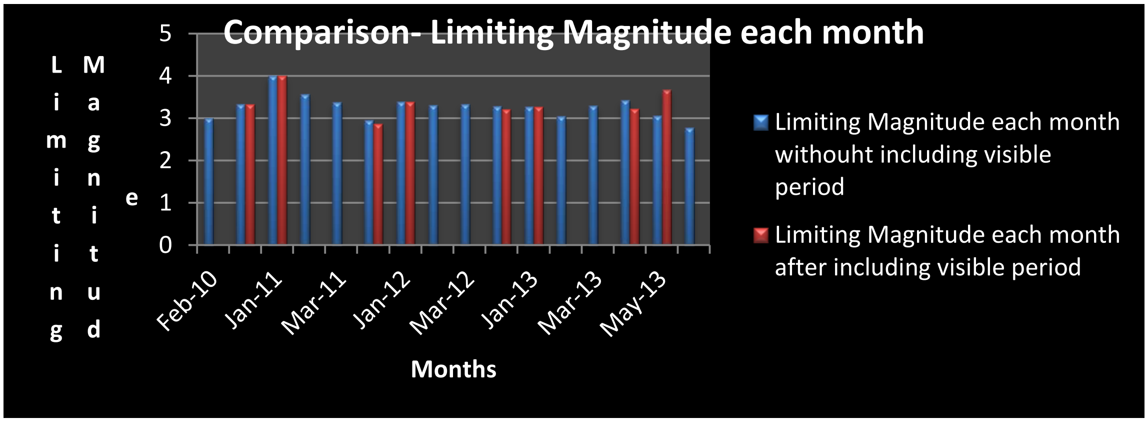
Analysis steps and insights:

Referring to the variables used in Proj4_File.py

- Row4 was used to store the **average limiting magnitude each year** without the use of data source 2 while row5 was used to store the average limiting magnitude each year along with the use of data source 2. The two results are compared
  Observation: *There is a wide range difference between the limiting magnitudes for year 2011*



- Row2 was used to store the **average limiting magnitude grouped by months each year** without the use of data source 2 while row3 was used to store the average limiting magnitude grouped my months each year along with the use of data source 2. The two results are matched
  <u>*Observation:*</u> *The results of row3 did not have output for all the months when compared with the the average limiting magnitude grouped my months each year along with the use of second data source*



- Row6 was used to find the mismatch between the constellation and visibility match. This provides a list of weak records that were used to calculate limiting magnitude .Row 8 was used to calculate the best visibility period for Orion, Leo and Crux. Row 7 was used to find the visible constellations in the month of February and March which were blank when average limiting magnitude was calculated using a combination of both data sources.

<u>*Observation:*</u>

- ➢ *Orion, Leo and Crux were not the best match to calculate limited magnitude in the month of February.*
- ➢ *Accuracy increases when limiting magnitude is calculated using Orion in January, Leo in April, Crux in May along with the list of visible constellations from row7 in February and March*

The following months are not the best visibility period for Crux, Leo and Orion but these wre used in observations

Crux	April	February	January	March
Leo	February	March	May	June
Orion	April	February	March	May

Crux, Leo and Orion are best visible in the following months

JANUARY	Eridanus	Reticulum	Taurus	Caelum	Camelopard	Dorado	Auriga	Mensa	Lepus	Orion
APRIL	Antlia	Leo	Sextans	Leo Minor	Chamaeleon	Ursa Major	south Crater			
MAY	Hydra	mid	Corvus	Centaurus	Coma Beren	Crux	Musca	Canes Vena	Virgo	

Following constellations can be used to calculate visibility since they are the brightest in February and March

FEBRUARY	Columba	Pictor	Canis Majo	Canis Minor	Monoceros	Gemini		
MARCH	Puppis	Volans	Lynx	Cancer	Pyxis	Carina	Hydra head	Vela

Things which didn't work: The dataset had record observations only for the first half of the year. As a result, there was a wide variation in the comparison of limiting magnitude each year. Data source three did not come to use as the data for the same was available in the second data source