Spatiotemporal analysis of flood losses

Nazia Ferdause Sodial

***Abstract*— As the climate continues to change, insurance serves as a platform for people to negotiate issues of loss [1]. Flood events are one of the most disastrous natural catastrophes and the losses could come in the path of the solvency of the insurance companies. In this study, the focus is to identify on a broader scale the regions of the USA which would require extra attention from the underwriters while writing flood policies. The timeline of 1981 – 2020 was considered to analyze the losses with respect to flood zones using different visual computational methods including machine learning algorithms as well.**

I. PROBLEM STATEMENT

Flood has been one of the most catastrophic disasters over the centuries causing huge financial losses. It is quite challenging for Underwriters to write the business in such a way that the company can minimize claims in the event of a flood disaster. The claims incurred by the insurance sector as a result of flood events in the United States, as well as the variables that contributed to these losses, will be explored in this study. FEMA manages the National Flood Insurance Program (NFIP), which is distributed to the public through a network of more than 50 insurance companies and NFIP Direct to reduce the socio-economic impact of floods. This study's main aim is to address the following questions:

1. Are there any patterns or seasonality of flood events?
2. Do coastal floods or storm surges cause more financial losses than river/inland flooding?
3. Is it possible for underwriters to target specific states or regions in order to reduce claims? Is there anything the insured can do to mitigate his or her or their losses?
4. Which areas are highly prone to flood disasters?

The dataset contains approximately 2.5 million instances of the flood losses of each location along with claims from 1973 to mid-2021. The dataset is best suited for this analysis because the NFIP has provided the data in a large high volume to construct an overview of flood events and spot potential flood risk management enhancements for the insured and insurer.

II. STATE OF THE ART

Kousky and Michel-Kerjan [2] examined a similar dataset (1978 – 2012) to identify six findings. Several hypotheses were tested about the nature and drivers of flood claims (e.g., the impact of flood zones, house characteristics, individual and collective mitigation, and repetitive loss properties) using fixed-effects regressions and other statistical analyses along with visual analysis, as well as uncovered quantitative relationships on the determinants of claims payments. The research also looks at the distribution of claims throughout time and space. Six hypotheses turned to be absolute findings in the paper[2]. The hypotheses are listed below:

1. Claims are more frequent in Special flood hazard areas (SFHAs) than outside SFHAs.

   Claims in V zones are higher than in A zones, which will be higher than outside SFHAs.

2. Claims as a percent of building value are higher for pre-FIRM claims than post-FIRM claims.

   The absolute magnitude of claims is higher for pre-FIRM claims.

3. Claims are larger for repetitive loss structures.
4. Elevated houses, houses with more than one floor, and houses with a basement have lower claims.
5. A higher degree of participation in the CRS by the community lowers the magnitude of flood claims, but it is unclear if this holds for the lowest score levels, which may not be associated with activities that substantially impact claim amounts.
6. Claims as a percent of building value have been declining over time, but total claims have been increasing.

These authors' hypotheses – 1 and 4 align with the research questions 2 and 3 of this paper and it will be computed using temporal heatmaps, spatial distribution maps, bar and stacked bar charts of Tableau along with a few basic statistics.

In Nan Wang's article [3], the focus was on exploring the flash flood disasters in China from 1950 to 2015. Clustering algorithm was used to highlight the distinct spatial and temporal patterns at different scales. The statistically significant clusters of flash flood disasters discovered in this study occur in specific places and last for a specific amount of time, which closely follows extreme rainfall patterns. Seasonal, annual, and even long-term persistent flash flood behaviours are all differentiated in the paper.

A similar approach will be followed in this paper. Density-based spatial clustering of applications with noise (DBSCAN) technique will be used in Python to identify spatiotemporal clusters of flood occurrences on the basis of two parameters the maximum radius and the maximum temporal duration.

III. PROPERTIES OF THE DATA

In this study, the dataset is derived from the NFIP system of record, staged in the NFIP reporting platform. All the locations of the USA were considered except Puerto Rico, Guam Islands, Alaska, Hawaii, and Honolulu as geographically they are quite far from the mainland which in turn would cause an issue in visualizing spatially as per the scope of this study. The dataset has 2,564,278 rows and 40 columns. The dataset contains a wide variety of information related to flood from 1973 to mid-2021. To reduce the size of the dataset, this study focuses on the timeline of 40 years, from 1981 to 2020. Some of the variables used in this study are the lat/long, state, the date and time of the loss, individual building and contents claims, flood depth, number of floors, post FIRM construction indicator, basement, and elevated building indicator. It also consists of Special flood hazard areas (SFHA) data which is very crucial in flood analysis.

A. Missing values

Most of the columns of the dataset have missing values. Missing values were removed from lat/long and flood zones. Missing values in building and contents claims columns can be interpreted that around 610112 building losses and 1491436 content losses caused by flood for those locations were below the deductible. Hence, no claim was paid by the insurer. These null values were replaced by zero for further analysis.

B. Outliers

As per figure 1(a), it is observed that the occurrence of floods in 2005, 2012, and 2017 years are exceptionally high. However, upon removing these peaks, in figure 1(b) 2008 and 2011 pose to be the outliers. This is because these are real events and cannot be considered as outliers. It is deduced that the years 2005, 2012, and 2017 had high occurrences of floods due to Hurricane Katrina, Sandy, and HIM (Harvey, Irma, and Maria). Similarly, there are certain claims in billions for building and claims in millions for contents but these are not outliers. Also, a lot of other factors like construction, occupancy, protection, exposure, etc are considered while writing the insurance. Hence, it is quite likely to have such exceptionally high claims.

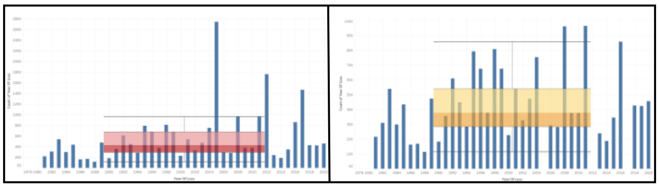


Figure 1. (a) Box and Whisker Plot on temporal distribution of flood occurrences for the timeline 1981-2020. (b) Box and Whisker on temporal distribution of flood occurrences excluding 2005, 2012, and 2017.

As per Figure 4, some outliers are observed spatially, as zone V must be along the coastline but few incidents are observed in the midwest of the USA. However, these outliers will be handled in clustering.

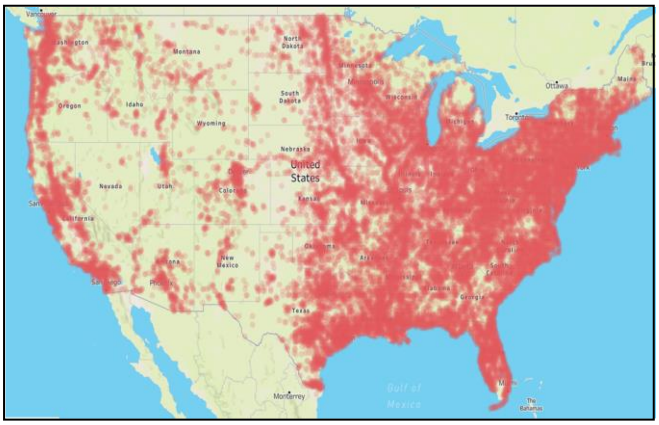


Figure 2. Spatial clusters and few scatterings of flood events over the USA

As per Figure 2, geographic clustering seems to be possible. The findings indicate that there is a spatial as well as a temporal pattern.

IV. Analysis

A. Approach

All the analysis will be done in Tableau except for Python was used for data preprocessing and DBSCAN. Before approaching the analysis, it is very critical to understand the flood zones. The area that will be swamped by a flood event that has a 1-percent chance of being equaled or exceeded in any given year is described as an SFHA. The base flood, often known as the 100-year flood, is a 1-percent yearly chance of flood. All the flood zones of groups A and V are the different types of SFHAs. The flood zones that belong to group A are caused due to river flood and group V due to storm surge or coastal flood. The FIRM also shows moderate flood danger zones, labeled Zone B or Zone X (shaded), which have 0.2-percent-annual-chance (or 500-year) flood. Zone C or Zone X (unshaded) are areas of little flood hazard, defined as places outside the SFHA [4].

Temporal Analysis Steps: Flood events occur every year and are usually associated with the monsoon season. However, this data can provide an insight into the seasonality of the losses caused by flood events. Tableau was used to visualize the temporal patterns.

- Visualize the date of loss data in a line plot to identify significant variations.
- Look for any monthly or yearly cyclic structure.
- Plot a heatmap of each month against all the years to observe any pattern.
- Investigate the reason behind such patterns.
- Look for variation in claim values of zone A and V against the months of each year using a heatmap.

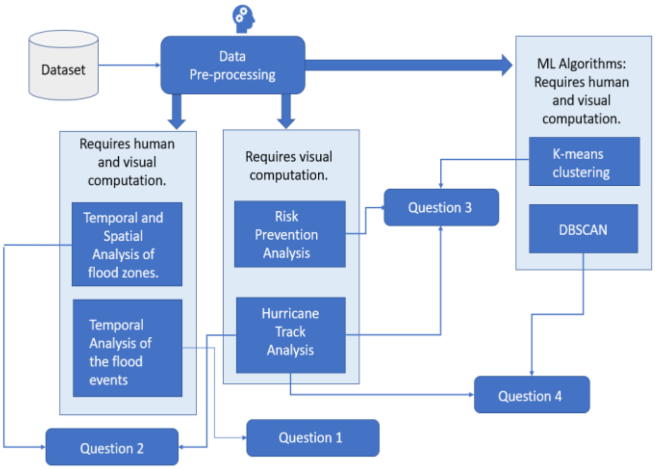

Dataset
Data Pre-processing
ML Algorithms: Requires human and visual computation.
K-means clustering
DBSCAN
Requires human and visual computation.
Temporal and Spatial Analysis of flood zones.
Temporal Analysis of the flood events
Requires visual computation.
Risk Prevention Analysis
Hurricane Track Analysis
Question 3
Question 4
Question 2
Question 1

Figure 3. Block Diagram of the approach taken to answer the research questions.

Spatial Analysis Steps: The hypothesis is that coastal areas or areas close to water bodies are likely to face flood disasters. Below steps will be followed:

- Visualize the density patterns overlayed over the map of the USA in Tableau.
- Look if certain regions have more concentration of flood losses along with flood zones and the possible explanation for such spatial patterns.

- Use K-means clustering to identify regions with high claims using Tableau.
- Use DBSCAN to identify Spatio-temporal clusters of flood events.

B. *Process*

Temporal Analysis: Also, due to rapid climate change, the temporal distribution has been changing within a year and over the years. Over this timeline, 2005 faced the highest number of flood events (~274k) and huge amount of claims (~17b) in the United States followed by 2012 and 2017. The heatmaps are divided into two maps of 20 years because the claims of the years 2005 and 2012 are so high that it fades the patterns of other years away. From the Figure 4, heatmaps of the years 1981 to 2020, it can be visually determined that there is no cyclic pattern of flood occurrences over a year. However, it can be inferred that the events are usually higher in three consecutive months which are August, September, and October. As per the Journal of meteorology, August is the month with the highest AMR and the highest monthly rainfall followed by the hurricane season. Slight temporal density is observed in May for a few years, as maximum monthly snowmelt-adjusted precipitation occurs, as snowmelt has a huge impact on streamflow in mountain catchments across the western United States.

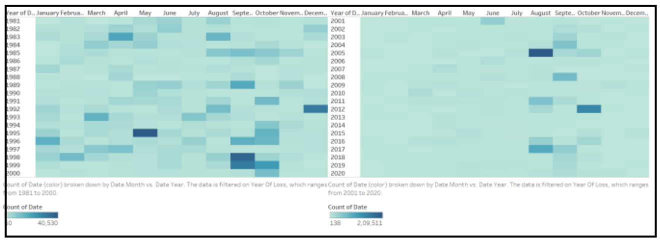
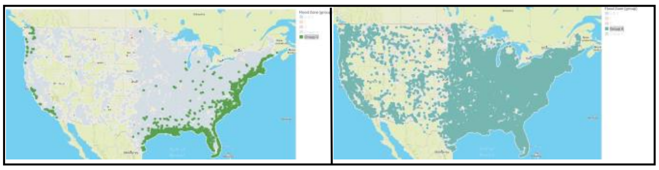


Figure 4. Heatmap of monthly flood events from 1981 to 2020.

Temporal and Spatial coherence of dominances of flood zones:

Over these 40 years, the total claims due to zone A is $49 billion and due to zone V, the claim is just $1.6 billion. However, as per the maps in Figure 5 (a) and (b), it is evident that the occurrence of river/inland floods is high compared to coastal floods and it is spread across the USA. In the Figure 5 (c) and (d), the blue shading depicts the lowest claims compared followed by white and then red. The shading of zone V is mostly white. So the heatmaps of three major flood zones, A (100-year flood zones), V (100-year flood zones), and B&X(500-year flood zones), reflect that the claims from zone V are mostly higher than any other zones. Compared to the ratio of occurrence, the claim of the coastal flood/storm surge is higher than the river/inland flood. This aligns with Kousky and Michel-Kerjan's [2] results. As per, The NOAA National Severe Storms Laboratory, storm surge is potentially deadly since it has the potential to flood vast regions along the shore. In coastal regions, extreme flooding can occur when storm surge coincides with normal high tide, resulting in storm tides of up to 20 feet or more in some situations. Hurricane Katrina (2005) is a perfect example of the havoc and damage that surges can wreak.



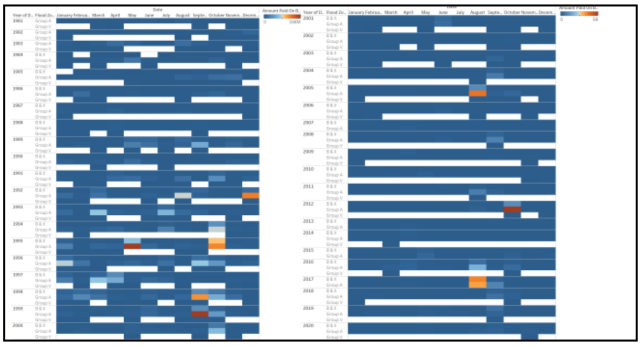


Figure 5. (a) Spatial distribution of flood events due to coastal flood/storm surge (b) due to river flood (c) heatmap of claims across months from 1981-2000 (d) heatmap of claims across months from 2001-2020

Spatial Analysis of Hurricane Tracks: Most people connect hurricanes with strong winds and the catastrophic damage they can cause, but precipitation-induced flood losses can be significant as well [5]. The USA is surrounded by coastal lines yet only the eastern regions face a high volume of flood events. After visualizing the hurricane tracks in NOAA, it is quite obvious that north-eastern and southern regions majorly face landfalls. As per Figure 6 (a) and (b), hurricanes played a very crucial role for 2005 and 2012's high claims. Whereas 6 (c) and (d) show no tracks of high category hurricanes over regions of the eastern USA, and hence, the claims of 2010 and 2014 were very low compared to the former. Although hurricanes mostly cause storm surge, Hurricane Harvey in 2017 revealed that, rather than wind or storm surges, the main cause of damage from a hurricane might be rainfall-induced inland flooding. Each incident adds to catastrophe modelling businesses' data archives, which are used to calibrate, validate, and enhance model capability [5].

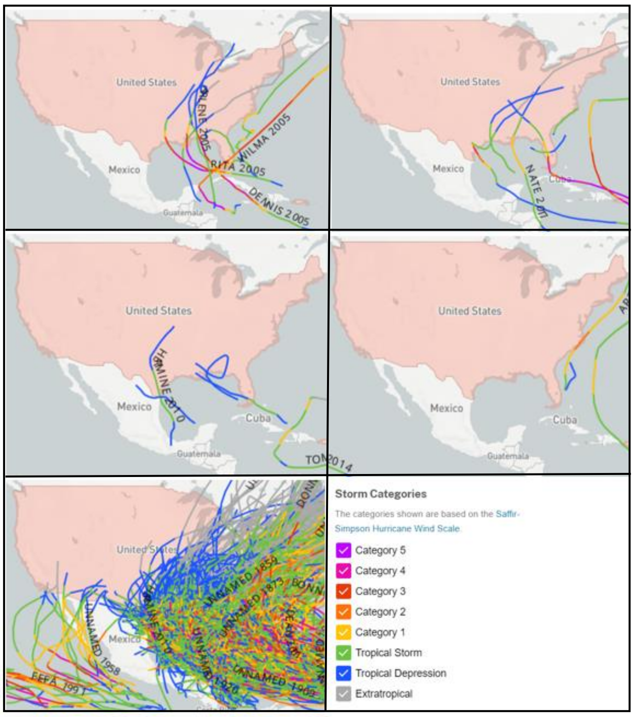


Figure 6. Hurricane tracks of (a) 2005 (b) 2012 (c) 2014 (d) 2010 (e) 1981-2020

K-means clustering: Pre-disaster planning and mitigation necessitate detailed spatial and temporal detail about flood hazards and their associated risks **[**6]. To understand the temporal and spatial pattern of claims of the USA from 1981 to 2020, series of choropleth maps were used as per Figure 7.

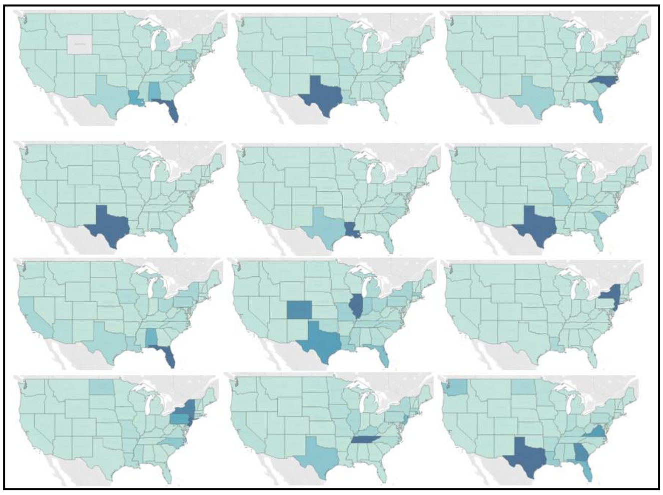


Figure 7. Claims distribution from 2020 to 2009 (top to right direction)

The southern part especially Texas seems to file high amount of cliams in the sample of temporal pattern.

To further cluster the regions on the basis of claims, K means clustering is used in Tableau. The data was normalized by Tableau and the variables were – building and content cliams. With Between-group Sum of Squares: 3.7552 and Within-group Sum of Squares: 0.046561, Tableau successfully created four clusters. Cluster 1 is created with 49 states, cluster 2 with 2 states, cluster 3 with 3 states, and cluster 4 with 2 states. From the value of the centroids, it seems that Texas and Louisiana filed similar and very high claims from 1981-2020, followed by Florida, New York, and New Jersey from cluster 3 and Mississippi and North Carolina from cluster 2.

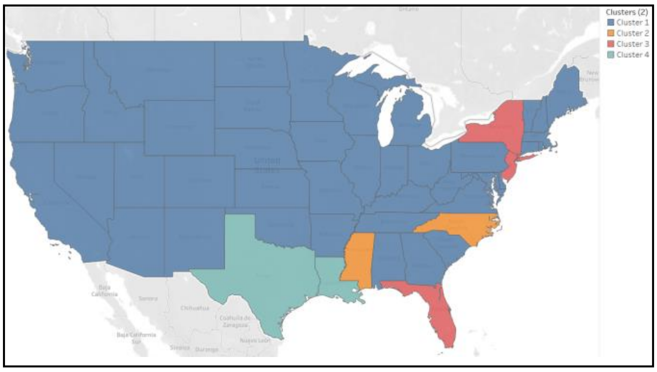


Figure 8. Clusters of states with respect to claims from 1981-202

Risk Prevention Analysis: The sight of flooding engulfing one's home appears to be a homeowner's worst fear. But being caught off guard and without insurance is even worse. Flooding alone caused $50 billion in property damage in the 1990s, accounting for more than two-thirds of all federal natural disasters [5]. In the same way, it is very important to understand the importance of risk prevention. In this section, it can be visually computed from Figure 9, that buildings that are elevated higher than base flood elevation with a greater number of floors face comparatively lower claims. Providing an accurate base flood elevation yields the greatest increment of benefits because it enables insurance premiums and building restrictions to be set based on a more realistic profile of where water will flow in the event of a flood [5]. Hence, having finished floor elevation more than base flood elevation can reduce premium as well as the claim. Additionally, having basements help reduce flood damage until it is not used to store stock. Furthermore, buildings built according to FIRM construction instructions face lower claims.

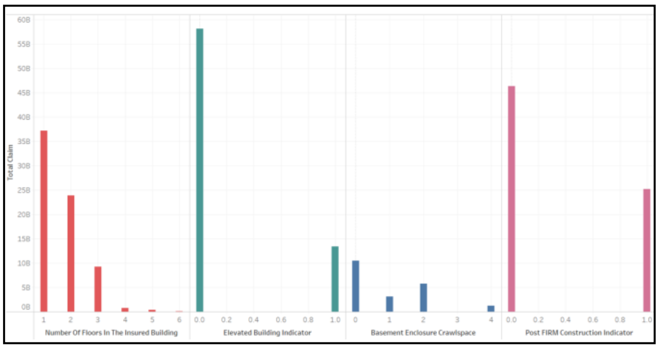
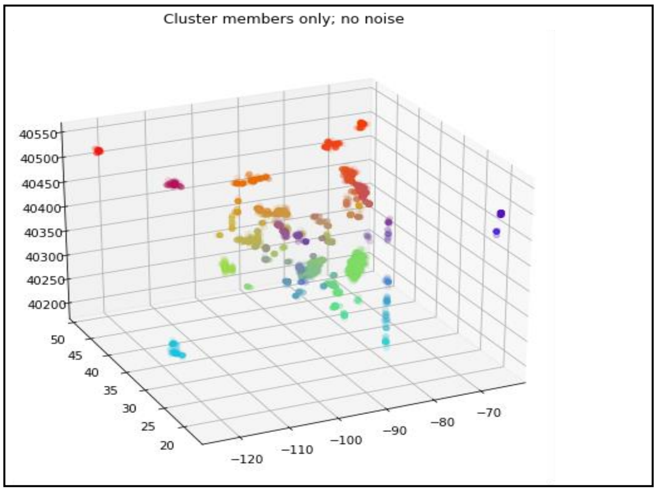


Figure 9. Bar plots of (a) number of floors (b) elevated building indicator (c) basement enclosure crawlspace (d) post FIRM construction indicator against total claims.

DBSCAN: As per National Weather Service, a single big rainfall event may be enough to create minor to major flooding in most cases. The biggest floods in the northeast, on the other hand, have typically been triggered by two huge storms occurring within a seven-day period. But not all flood events cause loss. In section, the focus is on identifying the regions with high flood events that cause loss.

In this section DBSCAN (density-based spatial clustering of applications with noise) algorithm is used. For this approach, 2010 data was considered as it was not influenced by hurricanes and the number of flood events is equivalent to the median of the temporal distribution. The lat/long were converted to radians. Similar to Nan Wang's [3] approach, the algorithm was iterated with several temporal distances and radial distances until clusters of flood events were dense and not loosely scattered. Finally, with a radial distance of 50 km per radian with 15 minpts and temporal distance as a week, 83 clusters were formed. These clusters can be visualized in Figure 10, a space-time cube without any noise.

The cluster results were further analysed with scatter plots in Tableau to visualize the spatiotemporal distribution of the flood events. Most of the clusters fall between July to September is a year. As per figure 10, spatially, the clusters are viewed in the northeast and south of the USA. Three clusters are observed in the west and a few in the midwest.



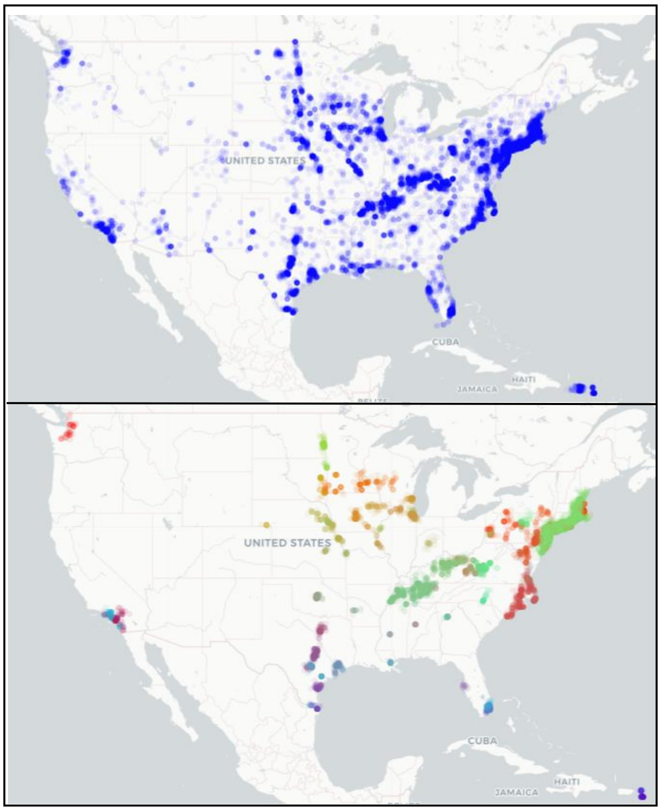


Figure 10. (a) Space-Time cube of flood events in the USA. (b) spatial distibution of flood events in 2010 (c) Clusters of flood loss events over the USA using DBSCAN

C. *Results*

To answer the first question, there are no strong patterns in the temporal distribution. But as per Figure 4, it can be inferred that flood loss events are most likely in the month of August, September, and October.

To answer the second question, coastal flood/storm surge may not occur as much as river flood events but they are highly disastrous due to high base flood elevation and several other causes due to their association with hurricanes and tides. Hence, the amount of building and contents claims are usually high for coastal flood/storm surge leading to being one of the most careful underwritings in the insurance sector.

To answer the third question, using K-means clustering the total claims are divided into four clusters of regions. Insured and the insurer must focus on the three clusters based on their claims to avoid losses by embracing risk prevention methods.

To answer the third question, the DBSCAN algorithm was used to form 88 Spatio-temporal clusters of dense flood loss events and they can be identified from Figure 10.

V. CRITICAL REFLECTION

In this study, the focus was on the occurrence of flood events that cause losses rather than just mere flood occurrence. Hence, it can be concluded that the occurrence of such flood events is most likely to happen in August, September, and October and the flood management policies must be in place.

However, as per the density based clustering of 2010, the occurrence is observed mostly from July to October. The relative infrequency of events and the abundance of sources of loss uncertainty, especially for extreme events, make achieving this key criterion of seasonality problematic in the case of natural hazards such as floods. The most appropriate way to deal with these concerns is to use a catastrophe model that creates a huge number of conceivable danger scenarios [6].

As per FEMA, just an inch of flood can cause up to $25,000 of damage. Flood insurance is a separate policy that covers buildings and contents in a building and it is essential for homeowners to protect their financial assets and at the same time, it is crucial for the insurance industry to write the business in such a manner that the claims do not come in the way of the solvency of the companies. The analysis of this study helped to identify that the northeast and few parts of the Midwest, south are prone to flood losses. It can be concluded that the west region of the USA faces very little flood calamity and does not require any attention while writing flood insurance policies. The clusters obtained with the help of DBSCAN for 2010, don't have many clusters in Louisiana whereas it files one of the most number and value of claims. One of the reasons is that in that year no major hurricane was faced by this state. However, to generalize the DBSCAN model, it must be run over a sample of a few more years to find a Spatiotemporal pattern over the years.

And the claims filed from those losses can depend on several factors like the flood protection system, contents type, catchment area, nearby dam disaster, etc. One of the challenges of this analysis was not being able to use base flood elevation data and the construction date of the buildings. The dataset had base flood elevations in thousands which are practically impossible as flood depth can never go so high. Similarly, a huge number of instances had future construction dates, which again questioned the reliability of those data. Hence, it is difficult to make any absolute conclusion based on the variables used from this dataset. However, certain different approaches could enhance the quality of this study.

The further scope of this study is to decompose complex temporal and spatial behaviour and get into the granular level of each season of a year. Also, the K-means algorithm can be further validated and the hyperparameters can be tuned and visually inferred to improve the accuracy of the model.

Table of word counts

Problem statement	244
State of the art	402
Properties of the data	462
Analysis: Approach	354
Analysis: Process	1321
Analysis: Results	174
Critical reflection	490

REFERENCES

[1] Elliott, R. 2021. *Underwater: Loss, Flood Insurance, and the Moral Economy of Climate Change in the United States*. New York

Chichester, West Sussex: Columbia University Press. https://doi.org/10.7312/elli19026.

[2] C. Kousky and E. Michel-Kerjan, “Examining Flood Insurance Claims in the United States: Six Key Findings: Examining Flood Insurance Claims in the United States,” *J. Risk Insur.*, vol. 84, no. 3, pp. 819–850, Sep. 2017, doi: 10.1111/jori.12106.

[3] N. Wang, L. Lombardo, M. Tonini, W. Cheng, L. Guo, and J. Xiong, “Spatiotemporal clustering of flash floods in a changing climate (China, 1950–2015),” *Nat. Hazards Earth Syst. Sci.*, vol. 21, no. 7, pp. 2109–2124, Jul. 2021, doi: 10.5194/nhess-21-2109-2021.

[4] FEMA glossary

[5] "Mapping the Zone Improving Flood Map Accuracy" report by The National Academy of Sciences.

[6] W. Mobley, A. Sebastian, R. Blessing, W. E. Highfield, L. Stearns, and S. D. Brody, “Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: a pilot study in southeast Texas,” *Nat. Hazards Earth Syst. Sci.*, vol. 21, no. 2, pp. 807–822, Mar. 2021, doi: 10.5194/nhess-21-807-2021