Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2017-278-RC2, 2017

Natural Hazards and Earth System Sciences
Open Access
Discussions

Interactive comment on "The relationship between precipitation and insurance data for flood damages in a region of the Mediterranean (Northeast Spain)" *by* Maria Cortès et al.

Anonymous Referee #2

Received and published: 18 September 2017

1 Summary of the article

The authors establish a linear relationship between the logarithm of the triggering rainfall and the logarithm of the resulting insurance claims of flash floods in the region of Catalonia (Spain). The significance and magnitude of the correlation coefficient is used as the main argument to proof the hypothesis that precipitation alone is sufficient to predict the magnitude of insurance claims resulting from flash floods in that very region.

2 Review

After this summary lets jump into the study itself. I will try to describe the work with my own words, stopping here and there to add my toughts and concerns.

2.1 Data

Three prinicipal sources of information are used by the authors. First the Inungama database from which basic data about flash floods in Catalonia are drawn. These are:

- affected municipalities
- affected basins
- start and end date of the event.

An event in the context of this article is therefore, as I understood it, an entry in the Inungama database.

At this point the authors know when a (flash) flood happend and which municpialies in which basin were affected. Now the second source of data is entering the stage: the flood damage data from the Spanish Insurance Compensation Consortium (CCS). The event data is at the municipality level and therefore the CCS is aggregated also at that level (which is the finest grain of spatial resolution for this study). By performing a join based on the smallest temporal distance between the event and the date of the insurance claim every event should now also have a variable called *Compensations*.

The last data source is meteorological data from the Spanish State Meteorological Agency. This should add another collumn to the data set with the accumulated 24h precipitation on the event day.

Suggestion 1: Because the focus of the article is on flash floods the authors should only include flash floods in their analysis. The difference between flash and non-flash floods must then be stated clearly maybe a (working) defintion of flash floods could be based on the length of the event (less than 24 hours).

Suggestion 2: Figure 4 is showing the number of floods and the amount of compensation per municipality. Some of the municipalities which have no flood event have compensation payments suggesting a flaw in the homogenisation procedure or simply a graphical one because the legend of figure 4b starts at 0 with a light pink tone.

2.2 Aggregation

If I got it straight the dataset should consit of entries with the following structure: a flash flood event i affected n_i municipalities in m_i basins. From the n_i affected municipalities $n_i - k_i$ received compensations of Y_i. The anticipated cause of the event is the 24h precipitation $X_{i,j}$ recorded at the day of the event at station j. Next the auhtors try to find the pair $(Y_i, X_{i,j})$ which yields the highest correlation in the log-log plane.

Let us, for the moment, assume that the hypothesis: payed compensation is a linear function of precipitation,

$$\log(Y) = a + b * \log(X)$$

is true. How could this be physically possible? First the compensation payed to cover the damage is caused by a flood event. The flood is produced by a stream (may it ephemeral or not) and this stream has a basin. Finnally the precipitation collected by the basin is the fuel for the catastrophic machinery producing the flood. It follows that only the amount of precipitation in the basin of the damage causing-stream should be related to the amount of compensation. Sounds logical tome. The authors find that the maximum precipitation over all affected basins has the highest correlation with the sum of compensations in the affected basins. This is a minor contradiction with the flow of reasoning presented which I assume also the authors used. But further problems

may emerge like that the damage itself depends on the number of damageable objects (exposure) in the basin aka at the time of the event. Let us assume that a rainfall of P_x is causing the total damage of a building in a basins of size A_x for all buildings with a distance to the stream of say d_x then only changing the number of building in the buffer d_x will result in considerable difference in the amount of compensation.

Suggestion 3: The exposure should be taking into account in other words a relative compensation should be formulated as the response variable in the analysis.

Suggestion 4: Adding a scatterplot of precipitation versus compensations for all used aggregation procedures would strongly enhance the understanding of the results.

2.3 Results

The authors present with figure 5 the key results of the regional analysis. Only guessing from the figure a linear model should be seriously influenced by the obersavtion at $x = -1$. If the log is the logarithm to the base 10 than this is a precipitation value of 0.1 mm which also seems unrealsitic. The authors also state a precipitation threshhold (100 mm) at which significant damages are observed suggesting that the probability of having a damage above 30.000 is maximized if the precipitation is above 100 mm no further explanation nor quantifcation is given.

Suggestion 5: Look at the observation with the low precipitation in more detail. Is it a measurement error? Maybe their is a wrong decimal sign? Is it really a flash flood and is it caused by precipitation? Generally the definition of the analysed data should be made more precise aka the obersavtions should be checked if they belong to the set of interest aka not comparing apples with oranges.

The analysis on the basin scale is focusing on a black and white example: a basin showing high correlation and therefore supporting the hypothesis of the authors and on the other hand a basin with low correlation contradicting the hypothesis (the mean

correlation for all basins is 0.47 (se +/- 0.4) which is rather low). To resolve the low correlation in the black basin the authors split the data set according to a population by maximizing the correlation coefficient turning the black into a white one.

Suggestion 6: Using the population as a basis for classifing rural and urban regions reminds me of using a dummy variable in regression from their it is only a slight jump to use population as variable in conjunction with preciptation. Using a ANOVA (or testing against a 0 slope of population or precipitation) would do the trick to see which one of the two is more important. But following suggestion 3 the influence of the population should vanish if and only if the hypothesis of a linear model in only influenced by precipitation is correct.

The last subset of observation is the MAB (metropolian area of barcelona) suggesting that a finer temporal grain (30 min) of the precipitation is enhancing the predtiction of compensation payments. Then the precipitation is correlated with the precipitation in 24h which results in a low correlation. Now the whole other data analysis is based on the 24h precipitation but the 30 min seems to be better suited. What are the implications for the 24h precipitation used for the other data sets?

Suggestion 7: Presenting scatter plots are much better suited than maps in my humble opinion. The whole point of the study is the assumption of linearity between precipitation and compensation and simple plot could demonstrade this with elegant ease.

3 Final Statement

I hope the review was not unpolite and has in any way offented the auhtors which was not at all my purpose. I think the study needs a major overhaul regarding the data preprocessing as well as the techniques used to draw conclusions.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2017-278, 2017.