Predicting Volcanic Eruptions: Case Study of Rare Events in Chaotic Systems with Delay

Justin Parra[1], Olac Fuentes[1],
Elizabeth Anthony[2], and Vladik Kreinovich[1]

Departments of [1]Computer Science and [2]Geological Sciences
University of Texas at El Paso, El Paso, TX 79968, USA,
jrparra2@miners.utep.edu, ofuentes@utep.edu,
eanthony@utep.edu, vladik@utep.edu

1. Outline

- Volcanic eruptions can be disastrous.
- It is therefore important to be able to predict them as accurately as possible.
- Theoretically, we can use the general machine learning techniques for such predictions.
- However, in general, such methods require an unrealistic amount of computation time.
- It is therefore desirable to look for additional information that would enable us to speed up computations.
- In this talk, we provide an empirical evidence that the volcanic system exhibit chaotic and delayed character.
- We also show how this can speed up computations.

2. Predictions Are Important

- Often, we want to predict future values $y(t_f)$ $(t_f > t_0)$ of different quantities y.
- To predict a future value, we can use the values of the related quantities $x_1(t), \ldots, x_n(t)$ for $t \le t_0$.
- For that, we need to know the dependence of $y(t_f)$ on the values $x(t) = (x_1(t), \ldots, x_n(t))$.
- In some practical situations, we know the desired dependence.
- For example, we know Newton's equations that describe the orbit of an asteroid.
- Thus, we can use these known equation to make the corresponding predictions.
- In other cases, however, we do not know the desired dependence.

3. Need for Machine Learning

- We can use the general techniques for determining the desired dependence from data.
- Such techniques are known machine learning.
- Examples: neural networks (in particular, deep learning networks), support vector machines, etc.
- To predict m steps into the future, we use patterns $(x, y(t_f))$, where:
  - $y(t_f)$ is the observed value y at moment t_f and
  - x is a collection of all the x-tuples $x(t)$ observed at moments $t \le t_f - m$.

4. We Face a Practical Challenge

- The machine learning computation time grows fast with the number of unknowns.
- As possible inputs, we have each of n values x_i measured at each of N_t moments of time.
- So, we need the dependence on $N_t \cdot n$ unknowns.
- When N_t is large, the number of unknowns is large.
- Thus, the corresponding computation require too much computation time.
- And indeed, successful predictions – e.g., using deep learning – require high-performance computers.
- To overcome this challenge, we need to limit moments of time used for training.

5. Such a Limitation Is Indeed Possible

- Suppose that we want to predict the weather in the next hour.
- The weather usually does not change during an hour.
- Thus, the most informative are current values $x(t_0)$.
- Knowing last year's weather will not help.
- To get predictions for the next day, it may be a good idea to also look for yesterday's weather.
- We will see if there is a tendency for the temperature to increase or to decrease.
- If we are currently in the Fall, then, to get predictions for the next summer:
  - today's data is probably useless,
  - it is much more useful to get data from last summer.

6. Case Study: Predicting Volcanic Eruptions

- An unexpected eruption can be a big disaster.
- The ancient city of Pompei was destroyed by a nearby volcano.
- The Cretan civilization was destroyed by a tsunami caused by a volcanic eruption.
- Nowadays, millions of people live in the close vicinity of active volcanos: Naples, Mexico City.
- This makes the task of predicting volcanic eruptions even more critical.

7. Specific Volcanoes

- We used the Aleutian chain of volcanoes that reaches from Alaska to Russia.
- Because of their location, their eruption affect major flight paths in the Pacific.
- As a result, they are heavily monitored, with seismic sensors near almost all of them.
- Of course, volcanos near Naples and Mexico City are heavily monitored too.
- However, these are solo volcanos, while there are about 30 Aleutian volcanos.
- Hence, Aleutian eruptions are more frequent, and we have more data to study.

8. What Information We Can Use to Predict Volcanic Eruptions

- When magma ascends to the surface, this massive movement causes some seismic activity.
- This causes ground deformation.
- Also, volcanic gases come out.
- Detecting deformations and gases requires complex on-site equipment, and all we get is a few numbers.
- In contrast, seismic waves can detected far way, and carry a lot of information.
- So, volcanic prediction techniques are based mostly on seismic activities.
- There exist techniques for predicting eruptions.
- However, these techniques are not perfect, more efficient and more accurate methods are needed.

9. Enter Delay and Chaos

- Delay means that inputs x_i affect y only after some time T_d (example: incubation period).
- Thus, to predict $y(t_f)$, we only need to consider $x_i(t)$ for $t \le t_f - T_d$.
- Chaos means even if we know the current state, we cannot predict the distant future.
- A small change in the initial conditions can lead to a drastic changes of the future.
- This is known as the *butterfly effect*.
- In precise terms, chaos means that the effect of x_i disappears after some time T_c.
- So, to predict $y(t_f)$, we only need to consider $x_i(t)$ for $t \ge t_f - T_c$.
- Thus, we only need values $x_i(t)$ for $t \in [t_f - T_c, t_f - T_d]$.

10. How to Detect Delay and Chaos Based on Data

- Delay and chaos means that:
  - for some m_0, $y(t_f)$ is not effected by $x_i(t_f - m_0)$,
  - for other m_0, $y(t_f)$ strongly depends on $x_i(t_f - m_0)$.
- So, to detect T_d and T_c, we need to find values m_0 for which $y(t)$ strongly depends on $q = x_i(t_f - m_0)$.
- Each q is uniquely determined by properties $q < q_0$ and $q \geq q_0$ for different q_0.
- So, in effect, we must find m_0 and q_0 for which $y(t_f)$ most depends on whether $x_i(t_f - m_0) < q_0$.
- For a discrete event like an eruption, we can build a decision three based on whether $x_i(t_f - m_0) < q_0$.
- Inequalities close to the top of the tree correspond to important inputs.

11. How to Detect Delay and Chaos (cont-d)

- One way to build a decision tree is to use an entropy method.
- Entropy S is the average number of binary questions needed to find the answer.
- If in p cases, we had positive answer, then entropy is

$$S = -p \cdot \log_2(p) - (1 - p) \cdot \log_2(1 - p).$$

- If y depends on $q < q_0$, then, when we take only cases for which $q < q_0$, the entropy decreases.
- E.g., if $q < q_0$ uniquely determines y, uncertainty disappears and entropy decreases to 0.
- As the top of the decision tree, we select m_0 and q_0 for which the average entropy decreases the most.

12. Let's Apply This to Volcanic Prediction

- We considered predictions for $t_f = t_0 + m$, with $m = 7$, 15, 20, and 180 days.
- To predict, we used the cumulative earthquake values $X_i(t_0 - m_0)$ for $m_0 = 7$, 15, 30, and 180.
- For each of these time periods m_0, we used two types of data:
  - the overall *number* of earthquakes in a certain zone during the period m_0, and
  - the *sum* of the magnitudes of all these quakes.
- For each type of data, we also used the differences between:
  - the average values over the given period and
  - the average values over the previous period.

13. Inputs (cont-d)

- This helps us gauge to what extent the seismic activity has intensified.
- Specifically, we used the following three differences:

$$\frac{X(t_0-7)}{7}-\frac{X(t_0-15)}{15};\quad \frac{X(t_0-15)}{15}-\frac{X(t_0-30)}{30};$$

$$\frac{X(t_0-30)}{30}-\frac{X(t_0-180)}{180}.$$

- So, for each zone, and for each the two data types, we use 7 different values:
  - 4 values corresponding to 4 time periods, and
  - 3 values corresponding to the 3 differences.
- Thus, for each zone, we considered $2\times 7=14$ values.

14. Inputs (cont-d)

- For each zone, we considered $2 \times 7 = 14$ values.
- The overall neighborhood of each volcano was divided into $3 \times 3 = 9$ zones:
  - by the distance to the volcano: 0–2.5 km, 2.5–5 km, and 5–15 km; and
  - by depth: 0–5 km, 5–15 km, and 15–30 km.
- For each of these 9 zones, we had 14 variables, so the overall number of variables was $9 \times 14 = 126$.

15. What We Expected and What We Observed

- Based on common sense, we expected that:
  - for predictions for $t_f \approx t_0$, the most important inputs are $X_i(t)$ with $t \approx t_0 - T_d$ ($t \approx t_0$ if no delay),
  - for $t_f \gg t_0$, the most important inputs are $X_i(t)$ with $t \approx t_0 - T_c$ ($t \ll t_0$ if no chaos).
- We considered 4 prediction problems – predictions for 7, 15, 30, and 180 days ahead.
- In all 4 cases, the most important input is $X_i(t_0 - 30)$ corresponding to:
  - the previous 30 days, and
  - the zone which is the closest to the volcano and the shallowest (distance 0–2.5 km and depth 0–5 km).

16. Discussion

- We expected to see two different values $t - T_d$ and $t - T_c$ – corresponding to delay and to chaos, depending on:
  - whether we want short-term predictions
  - or we want or long-term predictions.
- Surprisingly, we got the exact same value $T_c = T_d \approx 30$.
- So, for volcanic eruptions, the delay and the chaos periods are approximately the same.
- As a result, only values $X_i(t)$ with $t \approx t_0 - 30$ should be taken into account.
- More recent and more distant values $X_i(t)$ do not affect the prediction.

17. This Is Bad News and Good News

- It is bad news: we cannot predict more than 30 days into the future.
- It is good news:
  - by considering only values $X_i(t_f - 3)$,
  - we decrease number of inputs and
  - thus, we speed up machine learning.
- This is what we are working on right now; our preliminary results are promising.

18. Acknowledgments

This work was supported in part:

- by the National Science Foundation grants:
  - HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and
  - DUE-0926721, and
- by an award ‘from Prudential Foundation.