# Objective

Given a 100K catalog,

1. Replace each event in the last 90K with one in the first 10K years.
2. The replacement and the original events should be as close as possible.

Each event is characterized by a vector of losses in Subareas.

| | | Subarea | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | Country loss | State 1 loss | . | . | State S loss | County 1 loss | . | . | County C loss |
| 1 | 12 | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |

# Objective

Given a 100K catalog,

1. Replace each event in the last 90K with one in the first 10K years.
2. The replacement and the original events should be as close as possible.

Each event is characterized by a vector of losses in Subareas.

Similarity between two events is measured by a distance function of the two loss vectors.

For each event in the last 90K, use its nearest neighbor in the first 10K as the replacement.

| | | **Subarea** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | **Country loss** | **State 1 loss** | **.** | **.** | **State S loss** | **County 1 loss** | **.** | **.** | **County C loss** |
| 1 | 12 | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |

# Objective

Given a 100K catalog,

1. Replace each event in the last 90K with one in the first 10K years.
2. The replacement and the original events should be as close as possible.

Each event is characterized by a vector of losses in Subareas.

Similarity between two events is measured by a distance function of the two loss vectors.

For each event in the last 90K, use its nearest neighbor in the first 10K as the replacement.

The nearest neighbor (NN) selection is a secondary objective. The primary goal is to minimize differences in the new 100K's EPs and the "true" 100K's EPs. The NN selection is a stepping stone.

| | | **Subarea** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | **Country loss** | **State 1 loss** | . | . | **State S loss** | **County 1 loss** | . | . | **County C loss** |
| 1 | 12 | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |

# Objective

Given a 100K catalog,

1. Replace each event in the last 90K with one in the first 10K years.

2. The replacement and the original events should be as close as possible.

Each event is characterized by a vector of losses in Subareas.

Similarity between two events is measured by a distance function of the two loss vectors.

For each event in the last 90K, use its nearest neighbor in the first 10K as the replacement.

The nearest neighbor (NN) selection is a secondary objective. The primary goal is to minimize differences in the new 100K's EPs and the "true" 100K's EPs. The NN selection is a stepping stone.

- Even if the primary goal can be achieved in other approaches, we would still prefer the NN proxy. Possession of event level similarity has merits.

| | | **Subarea** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | **Country loss** | **State 1 loss** | . | . | **State S loss** | **County 1 loss** | . | . | **County C loss** |
| 1 | 12 | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |

# Objective

Given a 100K catalog,

1. Replace each event in the last 90K with one in the first 10K years.
2. The replacement and the original events should be as close as possible.

Each event is characterized by a vector of losses in Subareas.

Similarity between two events is measured by a distance function of the two loss vectors.

For each event in the last 90K, use its nearest neighbor in the first 10K as the replacement.

The nearest neighbor (NN) selection is a secondary objective. The primary goal is to minimize differences in the new 100K's EPs and the “true” 100K's EPs. The NN selection is a stepping stone.

- Even if the primary goal can be achieved in other approaches, we would still prefer the NN proxy. Possession of event level similarity has merits.

Previous work selects the nearest neighbor using L1 (Manhattan) distance function with modification.

| | | **Subarea** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | **Country loss** | **State 1 loss** | . | . | **State S loss** | **County 1 loss** | . | . | **County C loss** |
| 1 | 12 | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |

# Objective

Given a 100K catalog,

1. Replace each event in the last 90K with one in the first 10K years.
2. The replacement and the original events should be as close as possible.

Each event is characterized by a vector of losses in Subareas.

Similarity between two events is measured by a distance function of the two loss vectors.

For each event in the last 90K, use its nearest neighbor in the first 10K as the replacement.

The nearest neighbor (NN) selection is a secondary objective. The primary goal is to minimize differences in the new 100K's EPs and the "true" 100K's EPs. The NN selection is a stepping stone.

- Even if the primary goal can be achieved in other approaches, we would still prefer the NN proxy. Possession of event level similarity has merits.

Previous work selects the nearest neighbor using L1 (Manhattan) distance function with modification.

- Loss difference in each Subarea is multiplied by a factor. The factor seems adhoc and its motivation is unclear. The legacy document questioned its necessity.

| | | **Subarea** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | **Country loss** | **State 1 loss** | . | . | **State S loss** | **County 1 loss** | . | . | **County C loss** |
| 1 | 12 | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . |

# Evaluation metric

We evaluate the upsampled catalog by computing the difference between its EPs and the "true" catalog's EPs.

Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{s=1}^{S} \sum_{y=1}^{100\text{K}} \left[\text{loss}^{\text{upsampled}}\left(\text{Subarea}_s, \text{Year}_y\right) - \text{loss}^{\text{truth}}\left(\text{Subarea}_s, \text{Year}_y\right)\right]^2}{\sum_{s=1}^{S} \sum_{y=1}^{100\text{K}} \left[\overline{\text{loss}} - \text{loss}^{\text{truth}}\left(\text{Subarea}_s, \text{Year}_y\right)\right]^2}$$

where $\overline{\text{loss}} = \frac{1}{100\text{K} \cdot S} \sum_{s=1}^{S} \sum_{y=1}^{100\text{K}} \text{loss}^{\text{truth}}\left(\text{Subarea}_s, \text{Year}_y\right)$.

1. $\text{loss}\left(\text{Subarea}_s, \text{Year}_y\right)$ is the $y$-th ordered annual loss in $\text{Subarea}_s$.

| | Coefficient of Determination | | | |
|---|---|---|---|---|
| | **Overall** | **Country** | **States** | **Counties** |
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.991 | 0.992 | 0.990 | 0.980 |
| $L_1$ (Manhattan) | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_{0.25}$ | 0.957 | 0.947 | 0.957 | 0.949 |
| $L_{0.33}$ | 0.972 | 0.965 | 0.973 | 0.965 |
| $L_{0.4}$ | 0.978 | 0.973 | 0.979 | 0.971 |
| $L_{0.5}$ | 0.985 | 0.982 | 0.986 | 0.976 |
| $L_{0.6}$ | 0.987 | 0.985 | 0.988 | 0.977 |
| $L_{0.7}$ | 0.989 | 0.987 | 0.989 | 0.978 |
| $L_{0.75}$ | 0.989 | 0.988 | 0.989 | 0.978 |
| $L_{0.8}$ | 0.990 | 0.989 | 0.990 | 0.979 |
| $L_{0.9}$ | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_2$ with 2x weights on counties | 0.990 | 0.991 | 0.990 | 0.980 |
| $L_2$ with 5x weights on counties | 0.989 | 0.989 | 0.988 | 0.980 |
| $L_2$ with 10x weights on counties | 0.988 | 0.988 | 0.987 | 0.980 |
| $L_2$ with weights of 1/sqrt(SubArea total Loss) | 0.989 | 0.988 | 0.988 | 0.980 |
| Inner product | 0.310 | 0.341 | 0.189 | -0.254 |
| Symmetric cross entropy | 0.837 | 0.856 | 0.796 | 0.696 |

# Evaluation metric

We evaluate the upsampled catalog by computing the difference between its EPs and the "true" catalog's EPs.

Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{s=1}^{S} \sum_{y=1}^{100K} \left[ \text{loss}^{\text{upsampled}} \left( \text{Subarea}_s , \text{Year}_y \right) - \text{loss}^{\text{truth}} \left( \text{Subarea}_s , \text{Year}_y \right) \right]^2}{\sum_{s=1}^{S} \sum_{y=1}^{100K} \left[ \overline{\text{loss}} - \text{loss}^{\text{truth}} \left( \text{Subarea}_s , \text{Year}_y \right) \right]^2}$$

where $\overline{\text{loss}} = \frac{1}{100K \cdot S} \sum_{s=1}^{S} \sum_{y=1}^{100K} \text{loss}^{\text{truth}} \left( \text{Subarea}_s , \text{Year}_y \right)$.

1. $\text{loss} \left( \text{Subarea}_s , \text{Year}_y \right)$ is the $y$-th ordered annual loss in $\text{Subarea}_s$.

2. $r^2$ is the most commonly used evaluation for regression models. We shelve the exploration of evaluation metric for now. It can be a never-ending self debate.

| | Coefficient of Determination | | | |
|---|---|---|---|---|
| | **Overall** | **Country** | **States** | **Counties** |
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.991 | 0.992 | 0.990 | 0.980 |
| $L_1$ (Manhattan) | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_{0.25}$ | 0.957 | 0.947 | 0.957 | 0.949 |
| $L_{0.33}$ | 0.972 | 0.965 | 0.973 | 0.965 |
| $L_{0.4}$ | 0.978 | 0.973 | 0.979 | 0.971 |
| $L_{0.5}$ | 0.985 | 0.982 | 0.986 | 0.976 |
| $L_{0.6}$ | 0.987 | 0.985 | 0.988 | 0.977 |
| $L_{0.7}$ | 0.989 | 0.987 | 0.989 | 0.978 |
| $L_{0.75}$ | 0.989 | 0.988 | 0.989 | 0.978 |
| $L_{0.8}$ | 0.990 | 0.989 | 0.990 | 0.979 |
| $L_{0.9}$ | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_2$ with 2x weights on counties | 0.990 | 0.991 | 0.990 | 0.980 |
| $L_2$ with 5x weights on counties | 0.989 | 0.989 | 0.988 | 0.980 |
| $L_2$ with 10x weights on counties | 0.988 | 0.988 | 0.987 | 0.980 |
| $L_2$ with weights of 1/sqrt(SubArea total Loss) | 0.989 | 0.988 | 0.988 | 0.980 |
| Inner product | 0.310 | 0.341 | 0.189 | -0.254 |
| Symmetric cross entropy | 0.837 | 0.856 | 0.796 | 0.696 |

# Evaluation metric

We evaluate the upsampled catalog by computing the difference between its EPs and the "true" catalog's EPs.

Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{s=1}^{S}\sum_{y=1}^{100\text{K}}\left[\text{loss}^{\text{upsampled}}\left(\text{Subarea}_s, \text{Year}_y\right) - \text{loss}^{\text{truth}}\left(\text{Subarea}_s, \text{Year}_y\right)\right]^2}{\sum_{s=1}^{S}\sum_{y=1}^{100\text{K}}\left[\overline{\text{loss}} - \text{loss}^{\text{truth}}\left(\text{Subarea}_s, \text{Year}_y\right)\right]^2}$$

where $\overline{\text{loss}} = \frac{1}{100\text{K}\cdot S}\sum_{s=1}^{S}\sum_{y=1}^{100\text{K}} \text{loss}^{\text{truth}}\left(\text{Subarea}_s, \text{Year}_y\right)$.

1. $\text{loss}\left(\text{Subarea}_s, \text{Year}_y\right)$ is the $y$-th ordered annual loss in $\text{Subarea}_s$.
2. $r^2$ is the most commonly used evaluation for regression models. We shelve the exploration of evaluation metric for now. It can be a never-ending self debate.
3. Earthquake catalog downsampling uses a variant of $r^2$ but focuses on only a few order statistics.

| | Coefficient of Determination | | | |
|---|---|---|---|---|
| | **Overall** | **Country** | **States** | **Counties** |
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.991 | 0.992 | 0.990 | 0.980 |
| $L_1$ (Manhattan) | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_{0.25}$ | 0.957 | 0.947 | 0.957 | 0.949 |
| $L_{0.33}$ | 0.972 | 0.965 | 0.973 | 0.965 |
| $L_{0.4}$ | 0.978 | 0.973 | 0.979 | 0.971 |
| $L_{0.5}$ | 0.985 | 0.982 | 0.986 | 0.976 |
| $L_{0.6}$ | 0.987 | 0.985 | 0.988 | 0.977 |
| $L_{0.7}$ | 0.989 | 0.987 | 0.989 | 0.978 |
| $L_{0.75}$ | 0.989 | 0.988 | 0.989 | 0.978 |
| $L_{0.8}$ | 0.990 | 0.989 | 0.990 | 0.979 |
| $L_{0.9}$ | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_2$ with 2x weights on counties | 0.990 | 0.991 | 0.990 | 0.980 |
| $L_2$ with 5x weights on counties | 0.989 | 0.989 | 0.988 | 0.980 |
| $L_2$ with 10x weights on counties | 0.988 | 0.988 | 0.987 | 0.980 |
| $L_2$ with weights of 1/sqrt(SubArea total Loss) | 0.989 | 0.988 | 0.988 | 0.980 |
| Inner product | 0.310 | 0.341 | 0.189 | -0.254 |
| Symmetric cross entropy | 0.837 | 0.856 | 0.796 | 0.696 |

# Evaluation metric

We evaluate the upsampled catalog by computing the difference between its EPs and the "true" catalog's EPs.

Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{s=1}^{S} \sum_{y=1}^{100\text{K}} \left[ \text{loss}^{\text{upsampled}} \left( \text{Subarea}_s, \text{Year}_y \right) - \text{loss}^{\text{truth}} \left( \text{Subarea}_s, \text{Year}_y \right) \right]^2}{\sum_{s=1}^{S} \sum_{y=1}^{100\text{K}} \left[ \overline{\text{loss}} - \text{loss}^{\text{truth}} \left( \text{Subarea}_s, \text{Year}_y \right) \right]^2}$$

where $\overline{\text{loss}} = \frac{1}{100\text{K} \cdot S} \sum_{s=1}^{S} \sum_{y=1}^{100\text{K}} \text{loss}^{\text{truth}} \left( \text{Subarea}_s, \text{Year}_y \right)$.

1. $\text{loss} \left( \text{Subarea}_s, \text{Year}_y \right)$ is the $y$-th ordered annual loss in $\text{Subarea}_s$.
2. $r^2$ is the most commonly used evaluation for regression models. We shelve the exploration of evaluation metric for now. It can be a never-ending self debate.
3. Earthquake catalog downsampling uses a variant of $r^2$ but focuses on only a few order statistics.
4. Table shows Euclidean leads the board, but most other distance measures also yield sufficiently high $r^2$s.

| | Coefficient of Determination | | | |
|---|---|---|---|---|
| | **Overall** | **Country** | **States** | **Counties** |
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.991 | 0.992 | 0.990 | 0.980 |
| $L_1$ (Manhattan) | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_{0.25}$ | 0.957 | 0.947 | 0.957 | 0.949 |
| $L_{0.33}$ | 0.972 | 0.965 | 0.973 | 0.965 |
| $L_{0.4}$ | 0.978 | 0.973 | 0.979 | 0.971 |
| $L_{0.5}$ | 0.985 | 0.982 | 0.986 | 0.976 |
| $L_{0.6}$ | 0.987 | 0.985 | 0.988 | 0.977 |
| $L_{0.7}$ | 0.989 | 0.987 | 0.989 | 0.978 |
| $L_{0.75}$ | 0.989 | 0.988 | 0.989 | 0.978 |
| $L_{0.8}$ | 0.990 | 0.989 | 0.990 | 0.979 |
| $L_{0.9}$ | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_2$ with 2x weights on counties | 0.990 | 0.991 | 0.990 | 0.980 |
| $L_2$ with 5x weights on counties | 0.989 | 0.989 | 0.988 | 0.980 |
| $L_2$ with 10x weights on counties | 0.988 | 0.988 | 0.987 | 0.980 |
| $L_2$ with weights of 1/sqrt(SubArea total Loss) | 0.989 | 0.988 | 0.988 | 0.980 |
| Inner product | 0.310 | 0.341 | 0.189 | -0.254 |
| Symmetric cross entropy | 0.837 | 0.856 | 0.796 | 0.696 |

# Evaluation metric

We evaluate the upsampled catalog by computing the difference between its EPs and the "true" catalog's EPs.

Coefficient of determination:

$$r^2 = 1 - \frac{\sum_{s=1}^{S} \sum_{y=1}^{100K} \left[ \text{loss}^{\text{upsampled}} \left( \text{Subarea}_s, \text{Year}_y \right) - \text{loss}^{\text{truth}} \left( \text{Subarea}_s, \text{Year}_y \right) \right]^2}{\sum_{s=1}^{S} \sum_{y=1}^{100K} \left[ \overline{\text{loss}} - \text{loss}^{\text{truth}} \left( \text{Subarea}_s, \text{Year}_y \right) \right]^2}$$

where $\overline{\text{loss}} = \frac{1}{100K \cdot S} \sum_{s=1}^{S} \sum_{y=1}^{100K} \text{loss}^{\text{truth}} \left( \text{Subarea}_s, \text{Year}_y \right)$.

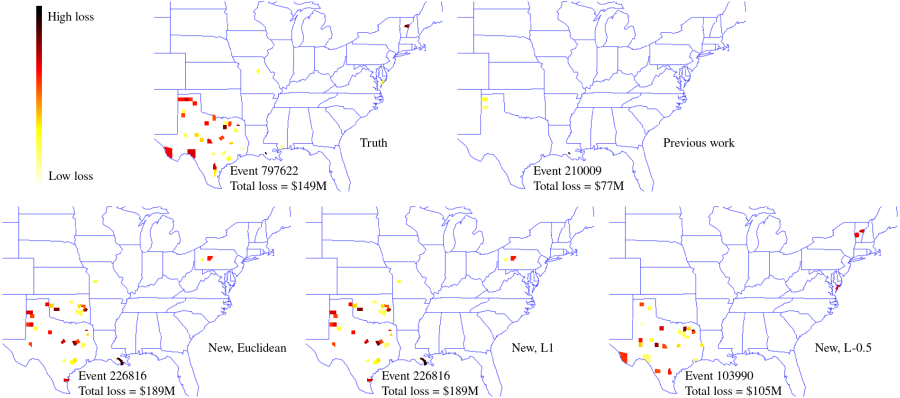
1. $\text{loss} \left( \text{Subarea}_s, \text{Year}_y \right)$ is the $y$-th ordered annual loss in $\text{Subarea}_s$.
2. $r^2$ is the most commonly used evaluation for regression models. We shelve the exploration of evaluation metric for now. It can be a never-ending self debate.
3. Earthquake catalog downsampling uses a variant of $r^2$ but focuses on only a few order statistics.
4. Table shows Euclidean leads the board, but most other distance measures also yield sufficiently high $r^2$s.
5. Computation shortcut: for each event, use the countrywide losses to determine its nearest 1000 neighbors. The computation is trivial because losses in 1-d space can be searched after sorting. Then among the 1000 candidates, compute distances and select the nearest neighbor $\Longrightarrow$ 10x speedup. Sparse representation of events $\Longrightarrow$ +10x speedup.

| | Coefficient of Determination | | | |
|---|---|---|---|---|
| | **Overall** | **Country** | **States** | **Counties** |
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.991 | 0.992 | 0.990 | 0.980 |
| $L_1$ (Manhattan) | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_{0.25}$ | 0.957 | 0.947 | 0.957 | 0.949 |
| $L_{0.33}$ | 0.972 | 0.965 | 0.973 | 0.965 |
| $L_{0.4}$ | 0.978 | 0.973 | 0.979 | 0.971 |
| $L_{0.5}$ | 0.985 | 0.982 | 0.986 | 0.976 |
| $L_{0.6}$ | 0.987 | 0.985 | 0.988 | 0.977 |
| $L_{0.7}$ | 0.989 | 0.987 | 0.989 | 0.978 |
| $L_{0.75}$ | 0.989 | 0.988 | 0.989 | 0.978 |
| $L_{0.8}$ | 0.990 | 0.989 | 0.990 | 0.979 |
| $L_{0.9}$ | 0.990 | 0.990 | 0.990 | 0.979 |
| $L_2$ with 2x weights on counties | 0.990 | 0.991 | 0.990 | 0.980 |
| $L_2$ with 5x weights on counties | 0.989 | 0.989 | 0.988 | 0.980 |
| $L_2$ with 10x weights on counties | 0.988 | 0.988 | 0.987 | 0.980 |
| $L_2$ with weights of 1/sqrt(SubArea total Loss) | 0.989 | 0.988 | 0.988 | 0.980 |
| Inner product | 0.310 | 0.341 | 0.189 | -0.254 |
| Symmetric cross entropy | 0.837 | 0.856 | 0.796 | 0.696 |

# 100 events sampled at random

Color scale is not linear to loss.

Similarity in spatial footprints alone does not dictate closeness between events.



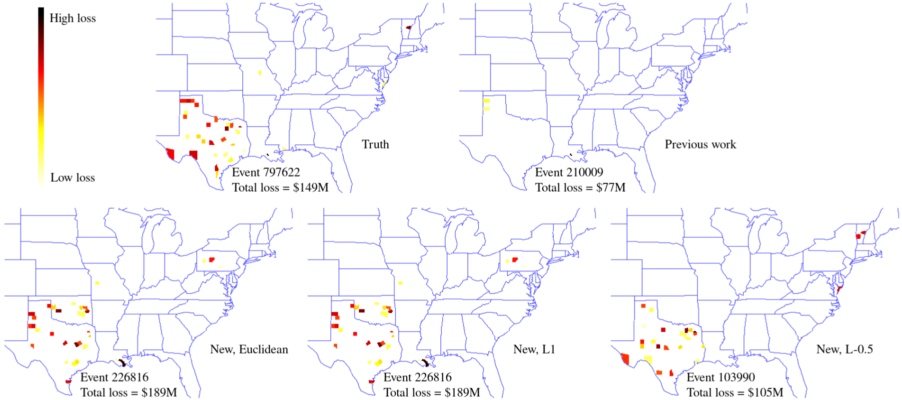
# 100 events sampled at random

Color scale is not linear to loss.

Similarity in spatial footprints alone does not dictate closeness between events.

Overall, New Euclidean preserves the spatial footprint slightly better, but has no consistent advantage over the previous work for every event.

We focus on Euclidean and $L_1$ moving forward.



# Exploration: feature engineering

| | | Subarea | | | | | | | | | Engineered features | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | Country loss | State 1 loss | . | . | State S loss | County 1 loss | . | . | County C loss | Mean loss of 8 closest counties to County 1 | . | . | Mean loss of 8 closest counties to County C | Mean loss of 16 closest counties to County 1 | . | . | Mean loss of 16 closest counties to County C | Mean loss of 32 closest counties to County 1 | . | . | Mean loss of 32 closest counties to County C | Mean loss of 64 closest counties to County 1 | . | . | Mean loss of 64 closest counties to County C |
| 1 | 12 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Using county losses alone to characterize events can be disastrous because of curse of dimensionality. Distance measure loses potency of distinguishing points in high-D space due to sparsity.

# Exploration: feature engineering

| | | Subarea | | | | | | | | | Engineered features | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | Country loss | State 1 loss | . | . | State S loss | County 1 loss | . | . | County C loss | Mean loss of 8 closest counties to County 1 | . | . | Mean loss of 8 closest counties to County C | Mean loss of 16 closest counties to County 1 | . | . | Mean loss of 16 closest counties to County C | Mean loss of 32 closest counties to County 1 | . | . | Mean loss of 32 closest counties to County C | Mean loss of 64 closest counties to County 1 | . | . | Mean loss of 64 closest counties to County C |
| 1 | 12 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Using county losses alone to characterize events can be disastrous because of curse of dimensionality. Distance measure loses potency of distinguishing points in high-D space due to sparsity.

Including Country and State losses in the characteristic vector provides more spatial context of an event, and is necessary to let the distance measure identify reasonable neighbors.

# Exploration: feature engineering

| | | Subarea | | | | | | | | | Engineered features | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | Country loss | State 1 loss | . | . | State S loss | County 1 loss | . | . | County C loss | Mean loss of 8 closest counties to County 1 | . | . | Mean loss of 8 closest counties to County C | Mean loss of 16 closest counties to County 1 | . | . | Mean loss of 16 closest counties to County C | Mean loss of 32 closest counties to County 1 | . | . | Mean loss of 32 closest counties to County C | Mean loss of 64 closest counties to County 1 | . | . | Mean loss of 64 closest counties to County C |
| 1 | 12 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Using county losses alone to characterize events can be disastrous because of curse of dimensionality. Distance measure loses potency of distinguishing points in high-D space due to sparsity.

Including Country and State losses in the characteristic vector provides more spatial context of an event, and is necessary to let the distance measure identify reasonable neighbors.

For building spatial context, a radical approach is to characterize the event using losses over a fine regular grid, and employ the feature engineering in convolutional neural net — convolve the loss image with kernels of various sizes, then concatenate the feature maps as the characteristic vector. However this could be too computationally heavy as the feature dimensionality can be massive.

As a compromise, we replace convolution of nearby pixels with averaging nearby county losses — the closest 8, 16, 32, 64 counties.

# Exploration: feature engineering

| | | Subarea | | | | | | | | | Engineered features | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Event | Country loss | State 1 loss | . | . | State S loss | County 1 loss | . | . | County C loss | Mean loss of 8 closest counties to County 1 | . | . | Mean loss of 8 closest counties to County C | Mean loss of 16 closest counties to County 1 | . | . | Mean loss of 16 closest counties to County C | Mean loss of 32 closest counties to County 1 | . | . | Mean loss of 32 closest counties to County C | Mean loss of 64 closest counties to County 1 | . | . | Mean loss of 64 closest counties to County C |
| 1 | 12 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 21 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 40 | . | | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | 42 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 61 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 62 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | 80 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Using county losses alone to characterize events can be disastrous because of curse of dimensionality. Distance measure loses potency of distinguishing points in high-D space due to sparsity.

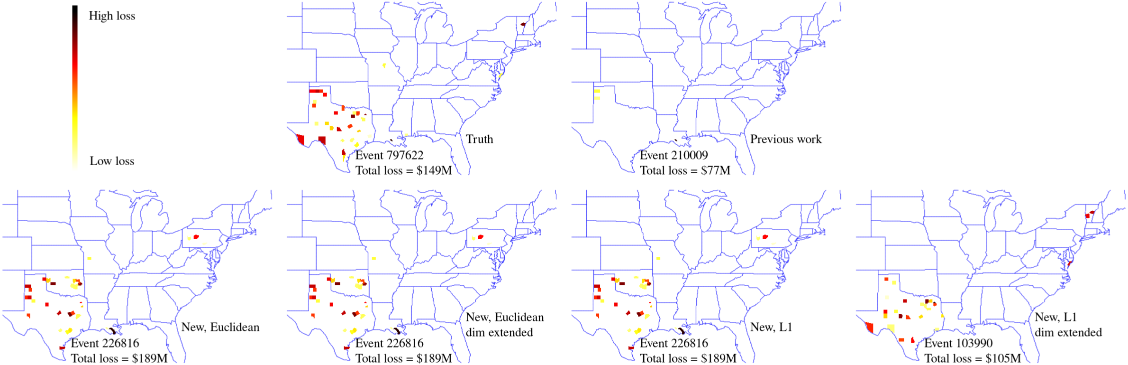
Including Country and State losses in the characteristic vector provides more spatial context of an event, and is necessary to let the distance measure identify reasonable neighbors.

For building spatial context, a radical approach is to characterize the event using losses over a fine regular grid, and employ the feature engineering in convolutional neural net — convolve the loss image with kernels of various sizes, then concatenate the feature maps as the characteristic vector. However this could be too computationally heavy as the feature dimensionality can be massive.

As a compromise, we replace convolution of nearby pixels with averaging nearby county losses — the closest 8, 16, 32, 64 counties.

With the engineered features, Country and State losses can be obsolete and may have negative impact. Keep them for now.

# 100 events sampled at random




Overall, "Euclidean + extended dimensions" preserves the spatial footprint slightly better, but still has no consistent advantage for all events.

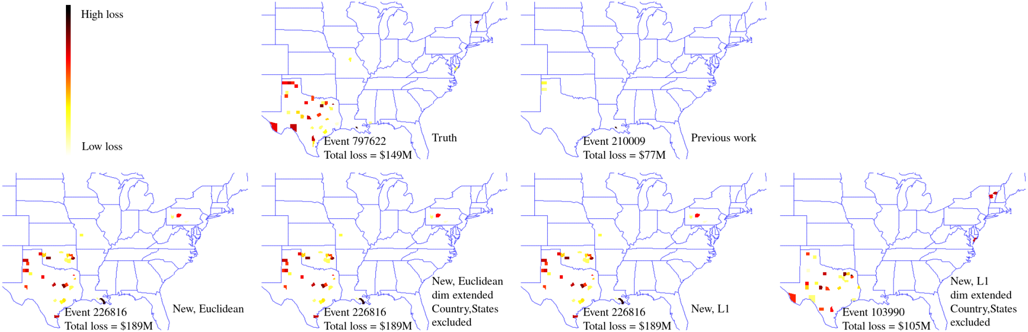
| | Overall | Country | States | Counties |
|---|---|---|---|---|
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.991 | 0.992 | 0.990 | 0.980 |
| $L_1$ (Manhattan) | 0.989 | 0.989 | 0.989 | 0.979 |

# Exploration: feature engineering, exclude country and states

| Year | Event | Country loss | State 1 loss | . | . | State S loss | County 1 loss | . | . | County C loss | Mean loss of 8 closest counties to County 1 | . | . | Mean loss of 8 closest counties to County C | Mean loss of 16 closest counties to County 1 | . | . | Mean loss of 16 closest counties to County C | Mean loss of 32 closest counties to County 1 | . | . | Mean loss of 32 closest counties to County C | Mean loss of 64 closest counties to County 1 | . | . | Mean loss of 64 closest counties to County C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 12 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | | | | |
| 2 | 21 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | | | | |
| 2 | 40 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | | | | |
| 2 | 42 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | | | | |
| 3 | 61 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | | | | |
| 3 | 62 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | | | | |
| 3 | 80 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | | | | |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

Subarea: Country loss … County C loss. Engineered features: Mean loss of 8 closest counties to County 1 … Mean loss of 64 closest counties to County C.

| | Overall | Country | States | Counties |
|---|---|---|---|---|
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.989 | 0.989 | 0.989 | 0.980 |
| $L_1$ (Manhattan) | 0.988 | 0.987 | 0.988 | 0.979 |

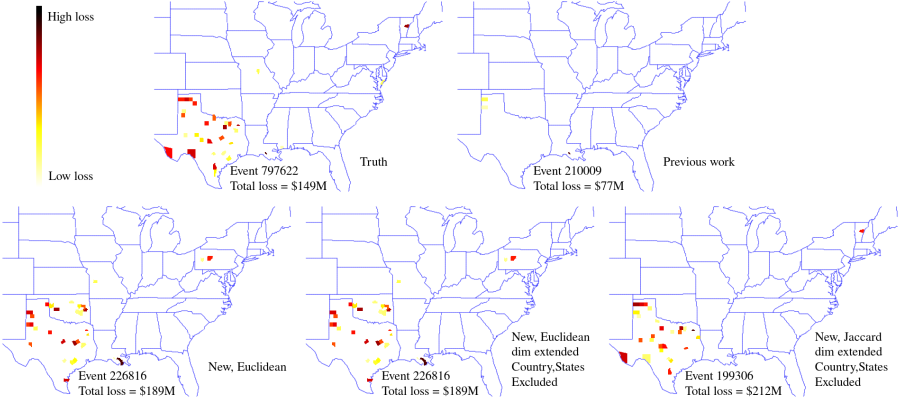



Overall, "Euclidean + dim extended + Country and States excluded" preserves spatial footprints better than "Euclidean + dim extended".

# Jaccard (binary distance)

| | Overall | Country | States | Counties |
|---|---|---|---|---|
| **Previous work:** | 0.988 | 0.991 | 0.986 | 0.974 |
| $L_2$ (Euclidean) | 0.989 | 0.989 | 0.989 | 0.980 |
| $L_1$ (Manhattan) | 0.988 | 0.987 | 0.988 | 0.979 |
| Jaccard | 0.930 | 0.928 | 0.916 | 0.897 |

Jaccard/binary is the best to preserve the spatial footprint, but it does not take loss magnitudes into account and thus yields poor $r^2$.



# Conclusion and next steps

1. Product team's approach appears good enough in terms of $r^2$. The scaling factors in their approach may be overengineered.

# Conclusion and next steps

1. Product team's approach appears good enough in terms of $r^2$. The scaling factors in their approach may be overengineered.
2. If they do not want to stay with their approach, we would recommend "Euclidean + feature engineering - country and states". We can facilitate friendly and fast software for the task (or are we taking over the project entirely?).

# Conclusion and next steps

1. Product team's approach appears good enough in terms of $r^2$. The scaling factors in their approach may be overengineered.
2. If they do not want to stay with their approach, we would recommend "Euclidean + feature engineering - country and states". We can facilitate friendly and fast software for the task (or are we taking over the project entirely?).
3. Sophistication: because Jaccard/binary is the best to preserve spatial footprint but struggles with $r^2$, we could try:
   3.1 Using Jaccard, store the $K$ (e.g. $K = 5$) nearest neighbors for each event.

# Conclusion and next steps

1. Product team's approach appears good enough in terms of $r^2$. The scaling factors in their approach may be overengineered.

2. If they do not want to stay with their approach, we would recommend "Euclidean + feature engineering - country and states". We can facilitate friendly and fast software for the task (or are we taking over the project entirely?).

3. Sophistication: because Jaccard/binary is the best to preserve spatial footprint but struggles with $r^2$, we could try:

   3.1 Using Jaccard, store the $K$ (e.g. $K = 5$) nearest neighbors for each event.

   3.2 For each event, if the Euclidean NN is among the Jaccard $K$ NNs, freeze the replacement event to the Euclidean NN.

   3.3 For all the other events, run stochastic optimization that maximizes $r^2$ by selecting one of the $K$ nearest neighbors.

# Paper topics ranked by preferences

1. Transformer for spatial model.

# Paper topics ranked by preferences

1. Transformer for spatial model.
2. Tsunami intensity acceleration using LASSO instead of convolutional neural net. Improve the approach and apply it to open source benchmarking for e.g. image regression/segmentation. Observe its competitiveness. Introduce the accelerator based on exploiting Cauchy inequality on the regularization path, and see if it is generic enough for all GLMs.

# Paper topics ranked by preferences

1. Transformer for spatial model.
2. Tsunami intensity acceleration using LASSO instead of convolutional neural net. Improve the approach and apply it to open source benchmarking for e.g. image regression/segmentation. Observe its competitiveness. Introduce the accelerator based on exploiting Cauchy inequality on the regularization path, and see if it is generic enough for all GLMs.
3. Catalog downsampling. The avenue will be a little different since the star of the show is the data structure designed for updating order statistics of high-D data. There are tons of algorithm researches on how to find/approximate quantiles in a data streaming environment. Our work could be relevant.

# Paper topics ranked by preferences

1. Transformer for spatial model.
2. Tsunami intensity acceleration using LASSO instead of convolutional neural net. Improve the approach and apply it to open source benchmarking for e.g. image regression/segmentation. Observe its competitiveness. Introduce the accelerator based on exploiting Cauchy inequality on the regularization path, and see if it is generic enough for all GLMs.
3. Catalog downsampling. The avenue will be a little different since the star of the show is the data structure designed for updating order statistics of high-D data. There are tons of algorithm researches on how to find/approximate quantiles in a data streaming environment. Our work could be relevant.
4. Four-point regriding. A long shot but there might be rich theoretical/numerical materials that can be mined or built.

**Temporary page!**

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, becau
LaTeX now knows how many pages to expect for this document.