Modelling and Classifying Stochastically Episodic Events

Colin Bellinger
Supervisor: Dr. B. John Oommen

Carleton University,
School of Computer Science

16 August, 2010

Overview

Stochastically Episodic (SE) Events

- Pattern Recognition (PR) of a *novel domain of problems*

Exemplary SE event problems include:

1. Large-scale tectonic movements
2. Industrial failure
3. Successful network attacks
4. Nuclear weapons tests

- Interwoven in a background of noise-like, non-SE events.

A Generalized PR Problem

- The previous problems fit into a novel category of PR
  - *SE event recognition*
- Identifiable via *four fundamental characteristics*
  1. Data: presents itself as a time-series
  2. State-of-nature: dominated by a single class
  3. The minority class: extremely challenging to identify
  4. SE events: rarely and randomly within the data sequence

A Domain Comparison

Traditional one-class PR versus SE event recognition

Domain	Temporal	ID Challenge	Imbalance *Type I*	Imbalance *Type II*	Interwoven
Mammogram	No	Low	Yes	Medium	No
Continuous typist recognition	No	Low	Yes	Medium	No
Password hardening	No	Low	Yes	Medium	No
Mechanical fault detection	No*	Low	Yes	Medium	No
Intrusion detection	No*	High	Yes	High	No
Oil spill	No*	High	Yes	Medium	No*
CTBT verification	Yes	High	Yes	High	Yes

- CTBT problem: exemplifies SE event domain
- Traditional one-class problems do not contain the collective set of SE event features
  - Possibility to reformulate some as SE event problems exists
  - Suggests a new point of exploration

The Challenge of Detecting SE Events

- As a PR problem, SE events pose *a significant challenge*
  - Extremely difficult to manually identify
    $\rightarrow$ "dirty" training and testing sets
  - Inherently rare
    $\rightarrow$ Inhibits the learning of a PDF
    $\rightarrow$ In practice, SE events are unpredictable
    $\rightarrow$ Insufficient to compile representative set for training/testing

Demonstration Domain

Comprehensive Nuclear Test-Ban-Treaty (CTBT)

- Bans the detonation of nuclear weapons
- Requires verification strategy

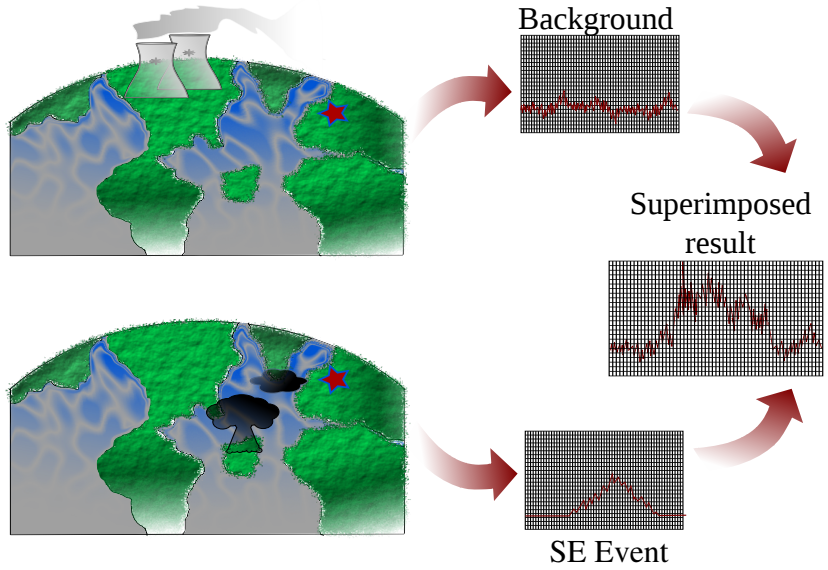
Proposed Strategy

- Measure four radionuclides
  - Emitted from industry and nuclear detonations
- Poses a PR problem
  - Distinguish bkgrnd source isotopes from detonation source

M&S: Problem Overview

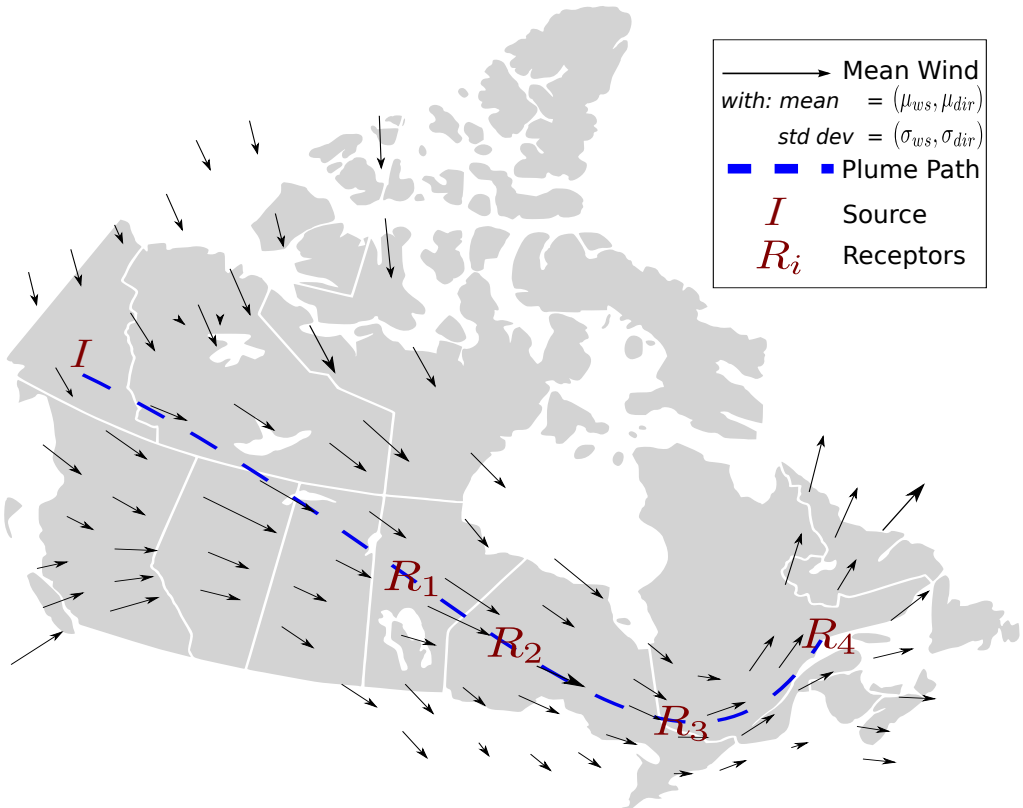
- *Necessity*
  - Data required for development of PR systems
  - "Real" SE event data: inherently difficult to acquire
- *M&S Challenge*
  - Majority of readings: bkgrnd noise
  - Bkgrnd dist'n: well-defined
  - SE event dist'n: extremely difficult to estimate
- *M&S Solution*: divide-and-conquer strategy
  - *Bkgrnd sources:*
    - Knowledge of propagation medium + bkgrnd emission rates
  - *SE event sources:*
    - Knowledge of propagation medium + probabilistic decisions

Divide-and-Conquer Strategy (Pictorial Overview)



1. Bkgrnd Module:
   - Simulate source's affect on receptor
   - Record in bkgrnd table
2. SE Event Generation Module:
   - Generate random SE events
   - Subjected to same meteorology
   - Record in SE event table
3. Merge tables and assign class labels

Background Simulation Setup

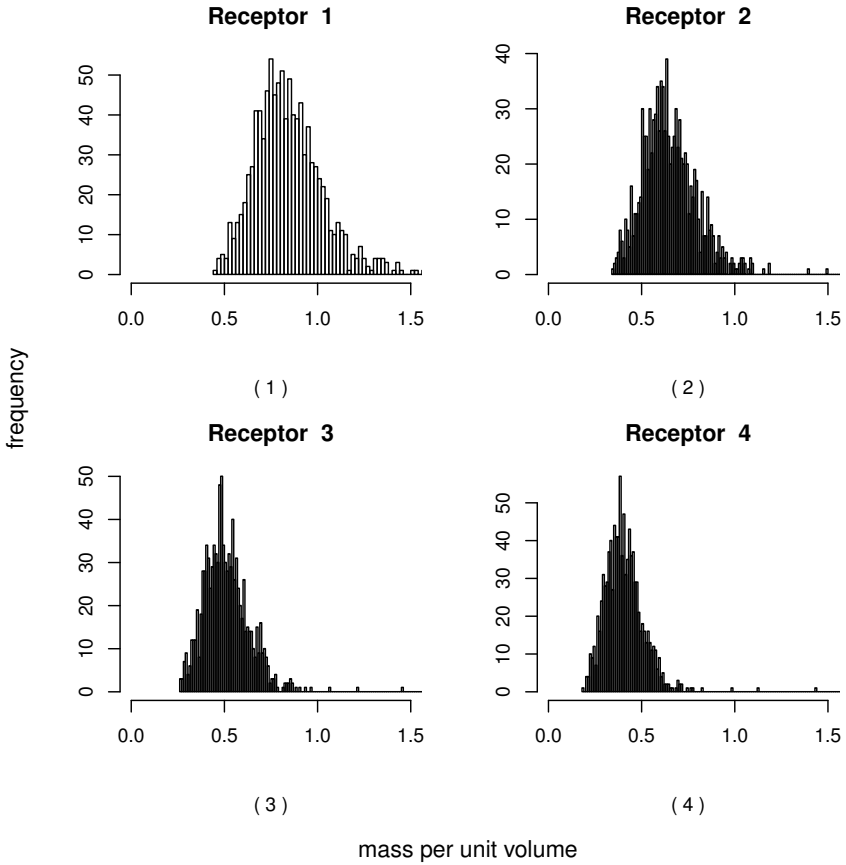


Simulation:

- Specify meteorological and emissions stats
- Modelled with continuous emission source Atmospheric Transport Model (ATM)

$$ATM = \chi_{GPlume} = \frac{Q}{2\pi\sigma_y\sigma_z\overline{u}} exp\left(-\frac{y^2}{2\sigma_y^2}\right)\left[exp\left(-\frac{(z-H)^2}{2\sigma_z^2}\right) + exp\left(-\frac{(z+H)^2}{2\sigma_z^2}\right)\right]$$

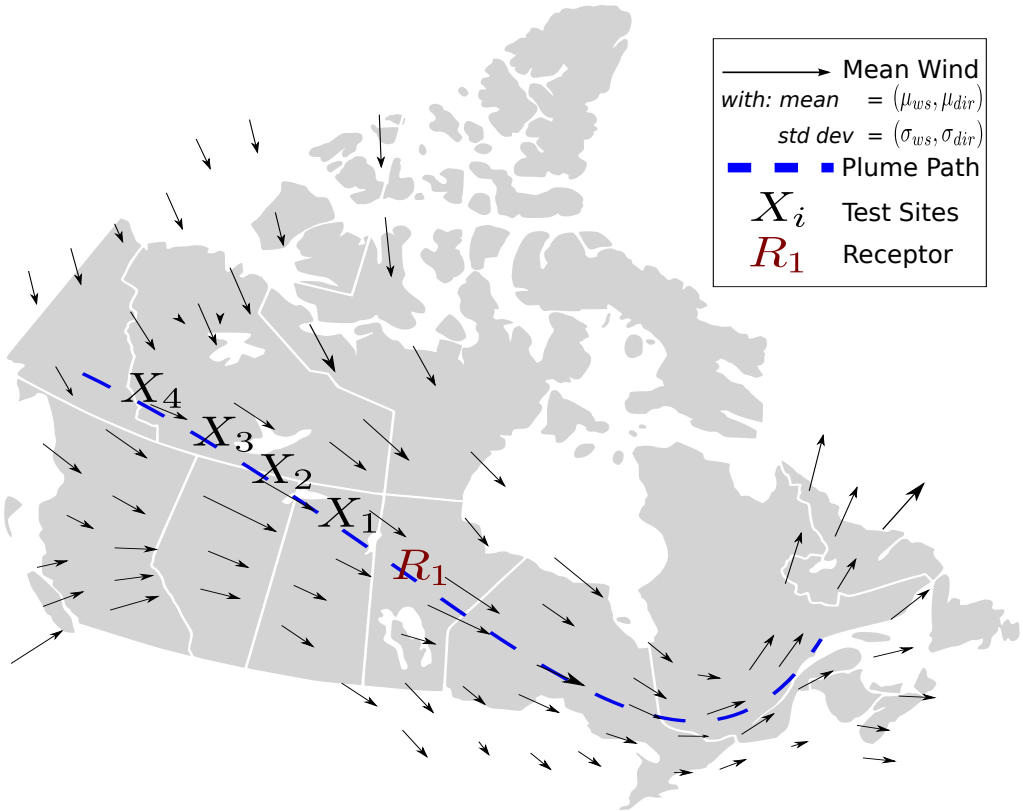
Background Simulation Results



Results:

- Hourly concentrations
- Distributions are well-behaved
- $\overline{\chi}$ decreases with successive R_i
- Decay overwhelms plume dispersion process
  - σ decrease with distances

SE Event Simulation Setup

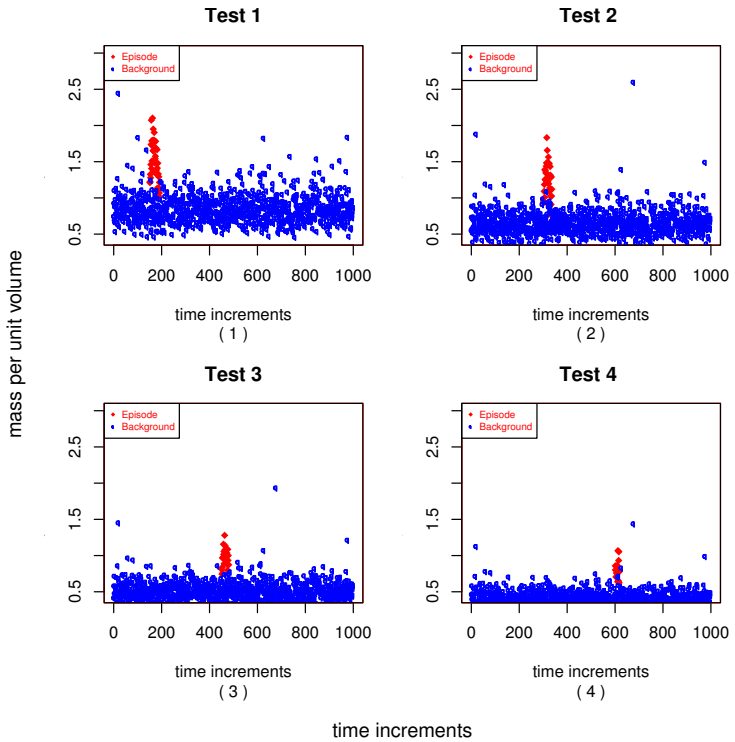


- Identical meteorological stats
- Demonstrate 4 upwind detonations
  - Illustrate effect on receptor site R_1
- Modelled with instantaneous point source ATM

$$ATM = \chi_{GPuff} = \frac{Q}{(4\pi t)^{\frac{3}{2}}(\sigma_x \sigma_y \sigma_z)^{\frac{1}{2}}} exp\left[- \frac{(x - \overline{u}t)^2}{2\sigma_x} + \frac{y^2}{2\sigma_y} + \frac{z^2}{2\sigma_z} \right]$$

Superimposed Simulation Results

- Smaller, narrower peaks
  - Cloud size and magnitude decrease with distance
- Visible PR challenges
  - Outliers in bkgrnd data
  - Seasonal met. variations
  - Classification based on single reading

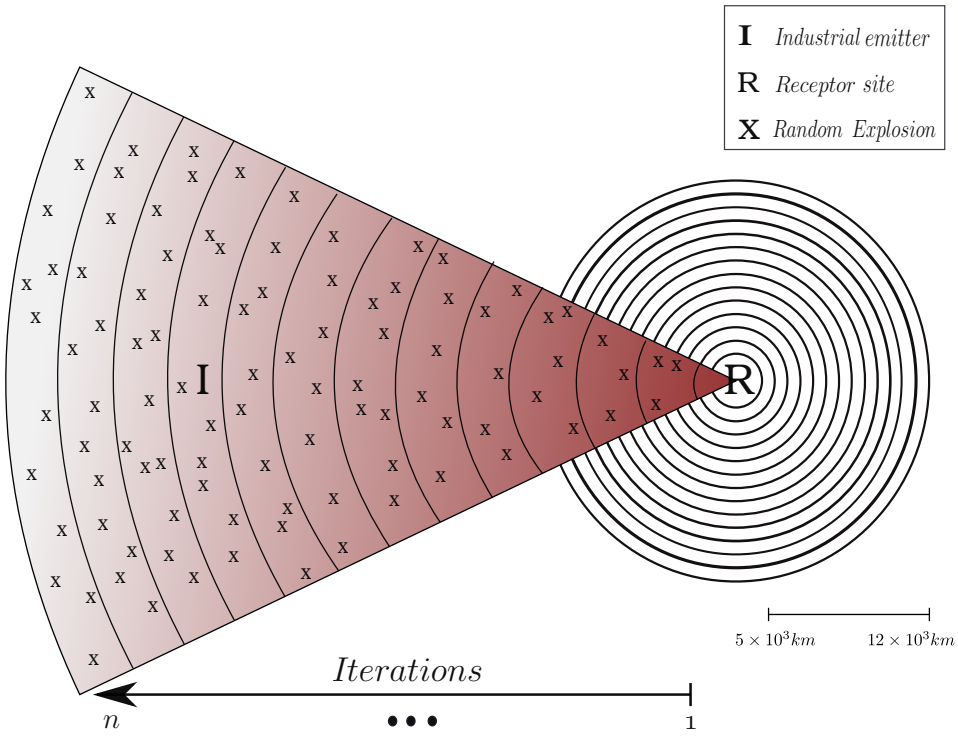


Classification: Scenario-Specific Approaches

- Possible scenarios:
  - S1: Small set of SE event data accurately labelled
  - S2: Insufficient training set
    - No appropriately labelled SE events
    - A few mislabelled as non-SE events
- Classification strategies
  - S1: Apply a standard binary/one-class classifier
    - Avoid overly optimistic performance estimates for binary classifiers
  - S2: Apply a one-class classifier
    - Insufficient data for binary classifiers
    - Noise due to mislabelled SE event $\rightarrow$ one-class classifiers applied in unconventional way

Classifier Assessment

- Results: ensemble mean AUC
  - Over ten iterations



- 23 detonation ranges
- 10 simulations per range
  - Random detonations
  - Constant bkgrnd and met stats
- Performance consideration
  - Overall
  - Function of distance
  - Expanded feature-space

Overall Performance

Scenario 1

	Mean	Max	Min	STDV
NB	0.772	0.939	0.504	0.074
MLP	**0.869**	0.976	**0.674**	0.067
NN	0.741	0.913	0.584	0.071
J48	0.774	**0.98**	0.500	0.148
SVM	0.528	0.813	0.500	0.065
ocNN	0.540	0.875	0.496	0.087
PDEN	0.487	0.943	0.182	0.156
socNN	0.603	0.842	0.405	0.094
AA	**0.656**	**0.970**	**0.251**	0.140

- ANN classifiers are superior
  - MLP (binary)
    - Low variability
  - AA (one-class)
    - Considerable variability

Overall Performance

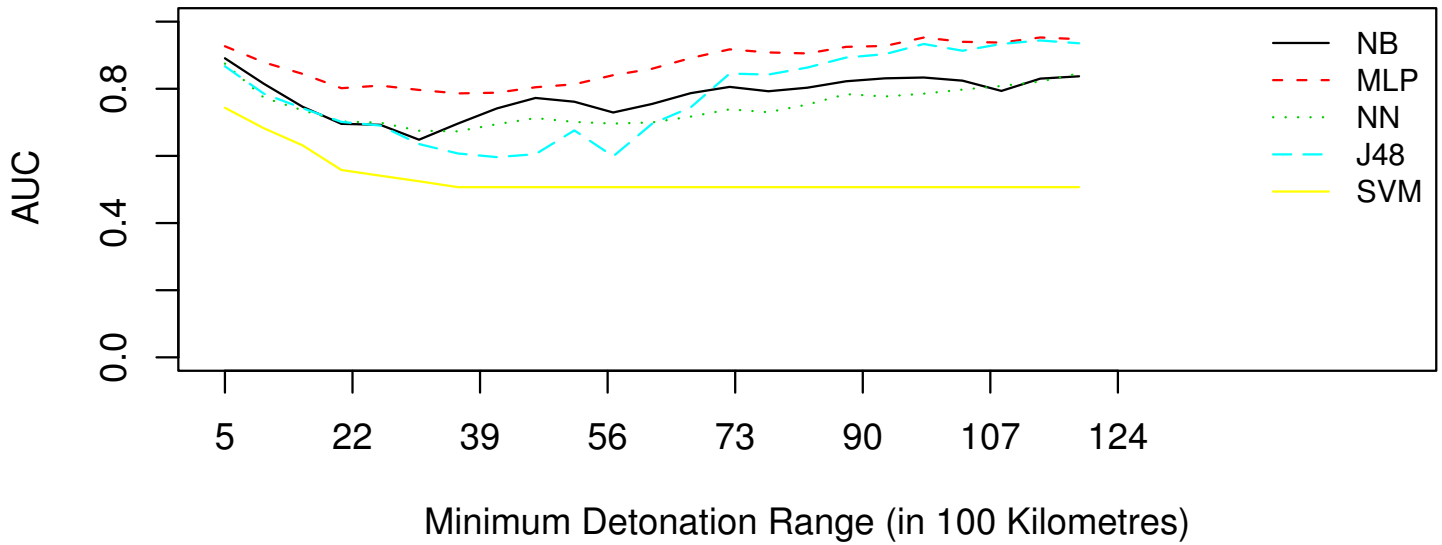
Scenario 2

	Mean	Max	Min	STDV
ocNN	0.505	1	**0.496**	0.042
PDEN	0.507	1	0.075	0.185
socNN	0.587	1	0.292	0.171
AA	**0.621**	1	0.024	0.225

- Significantly more challenging
  - On average: top classifiers degrade only slightly
  - Minimum AUC extremely poor

Performance as a Function of Distance

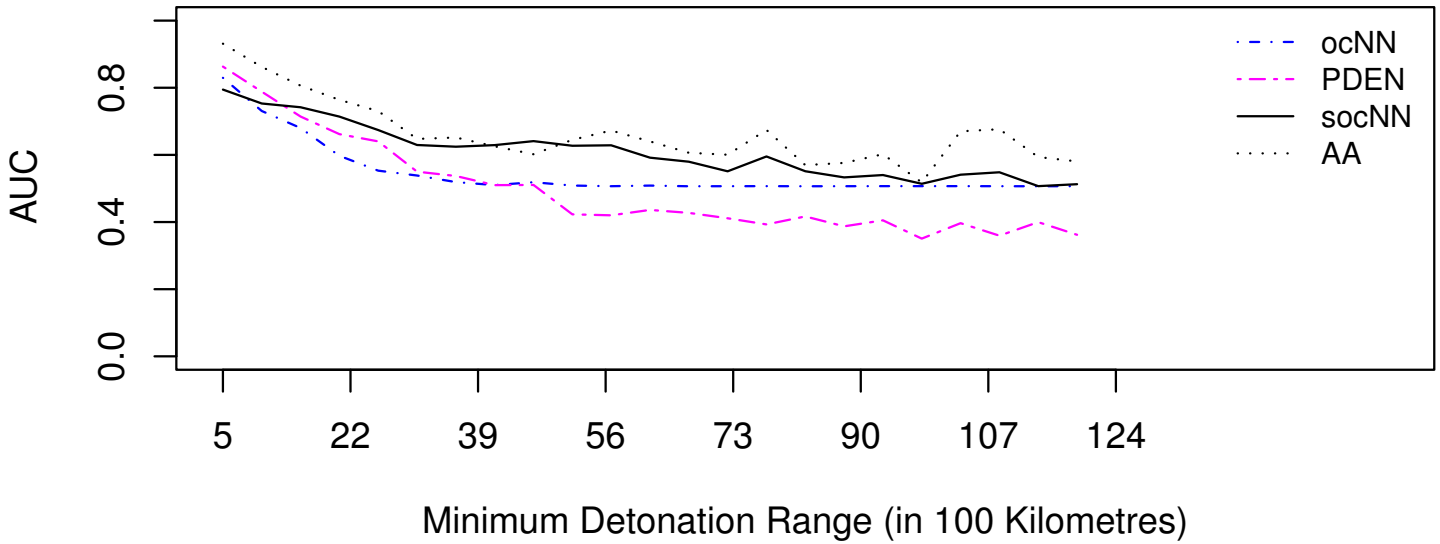
Scenario 1: Binary Learners



- MLP superior at all distances
- NB, NN and J48: comparable
- Hull coincides with industrial source

Performance as a Function of Distance

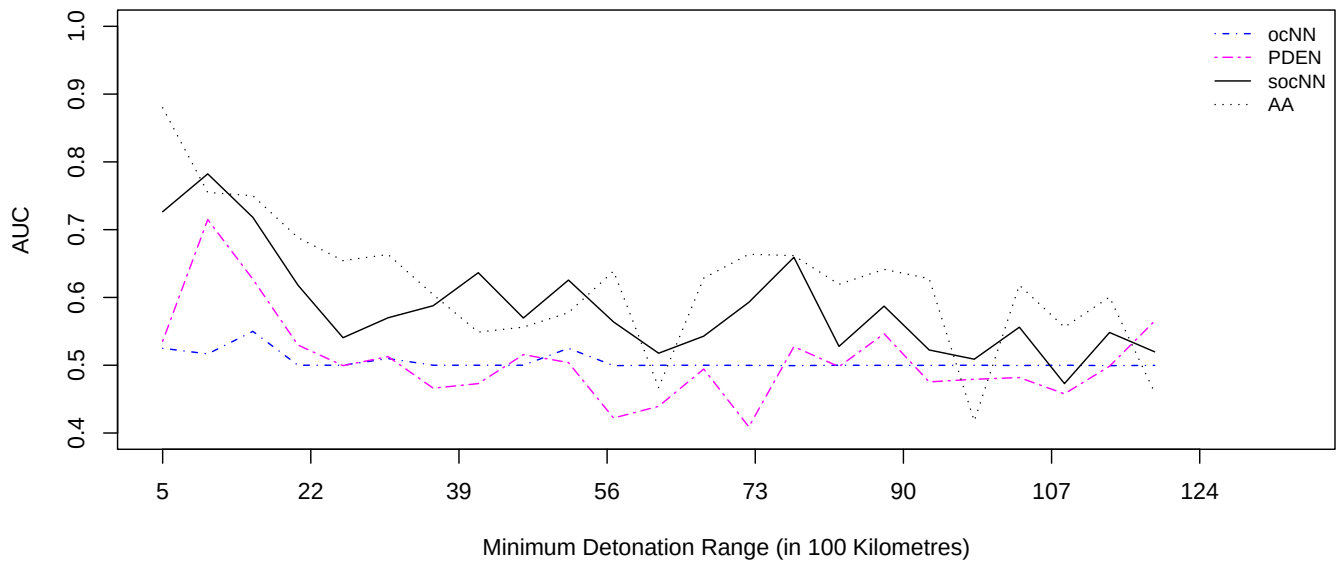
Scenario 1: One-class Learners



- AA and socNN generally superior
- AUC degrades with detonation distances
- Unable to associate low values with detonations

Performance as a Function of Distance

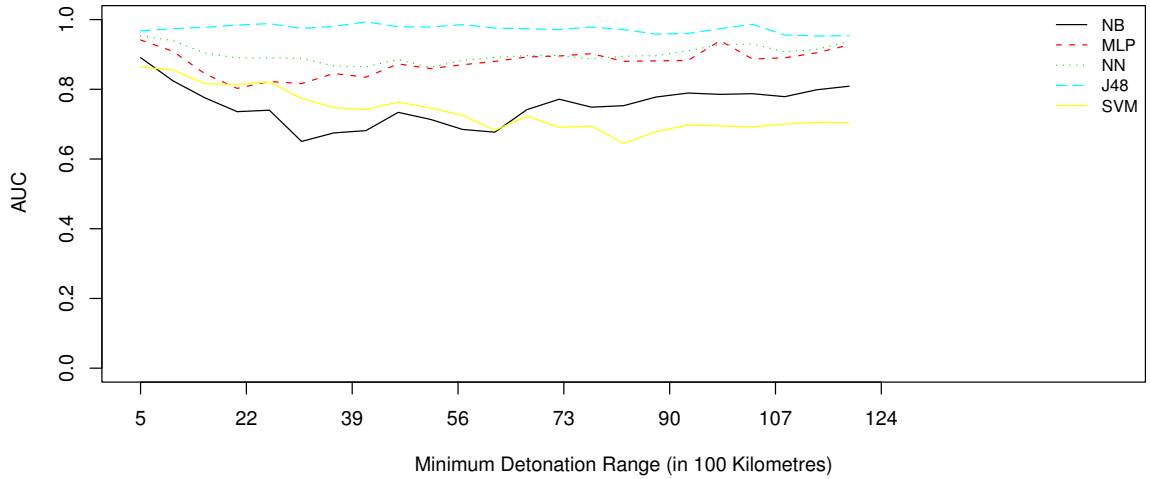
Scenario 2



- AA and socNN: similar to S1
- Highly variable
  - Nature of SE event in training set has significant affect

Expanded Feature-Space

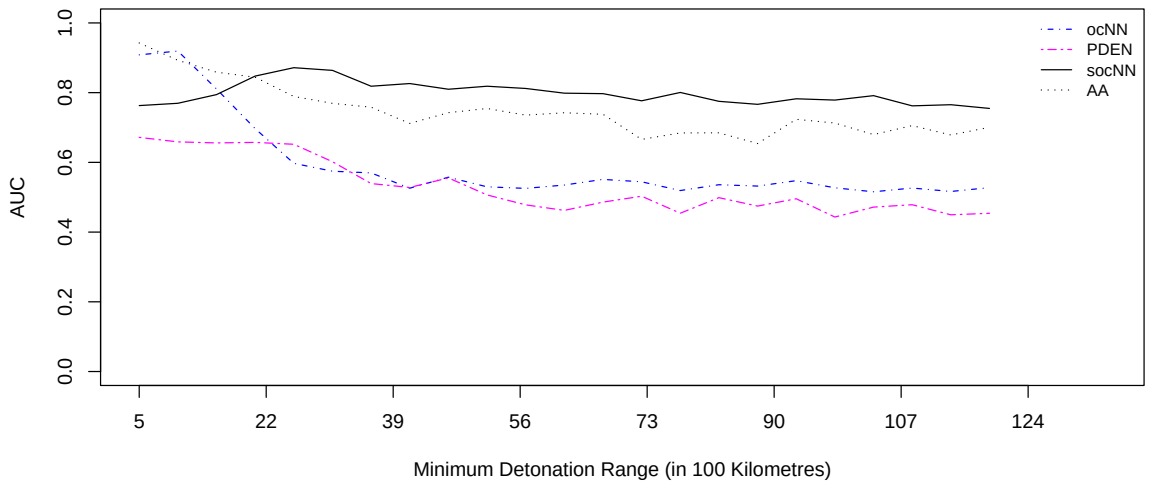
Scenario 1: Binary Learners



- WD: classifiers learn direction to bkgrnd source
- All classifiers benefit
  - Top 3: J48, NN, MLP
  - Hull significantly reduced
- J48: nearly perfect

Expanded Feature-Space

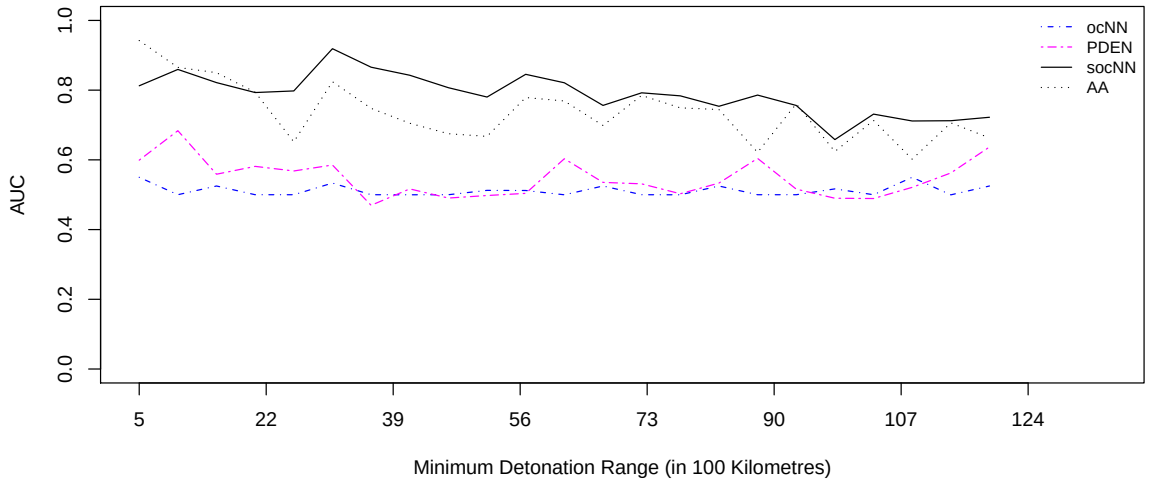
Scenario 1: One-class Learners



- All classifiers benefit
- socNN improved more than AA
- Similar to S1 binary
  - NN approach improved more than ANN approach

Expanded Feature-Space

Scenario 2



- Performances increase and stabilize
- socNN improves more that AA
  - socNN generally superior

Conclusion

Contributions

1. Modelling and Simulation
   - Framework for exploration of SE events
     - Propagated through background of noise-like, non-SE events
   - Applications:
     - PR system development + exploratory M&S

2. Pattern Recognition
   - New category of PR problems
     - SE event recognition
   - Two SE event classification techniques

Conclusion

Future Work

- Modelling and Simulation
  - Explore:
    - Environmental and industrial disaster scenarios
  - Incorporate:
    - Increasingly sophisticated ATM into framework
- Pattern Recognition
  - Emphasize temporal nature of the data
    - Explore PR as an early warning system
  - Derive:
    - Increasingly sophisticated SE event recognition algorithms