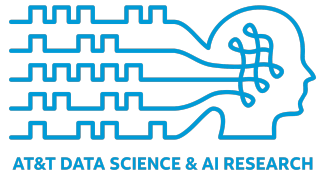
Kernel Mean Embedding Based Hypothesis Tests for Comparing Spatial Point Patterns

Raif Rustamov and James Klosowski
Data Science and AI Research
AT&T Labs

June 5, 2020



AT&T

Motivating example

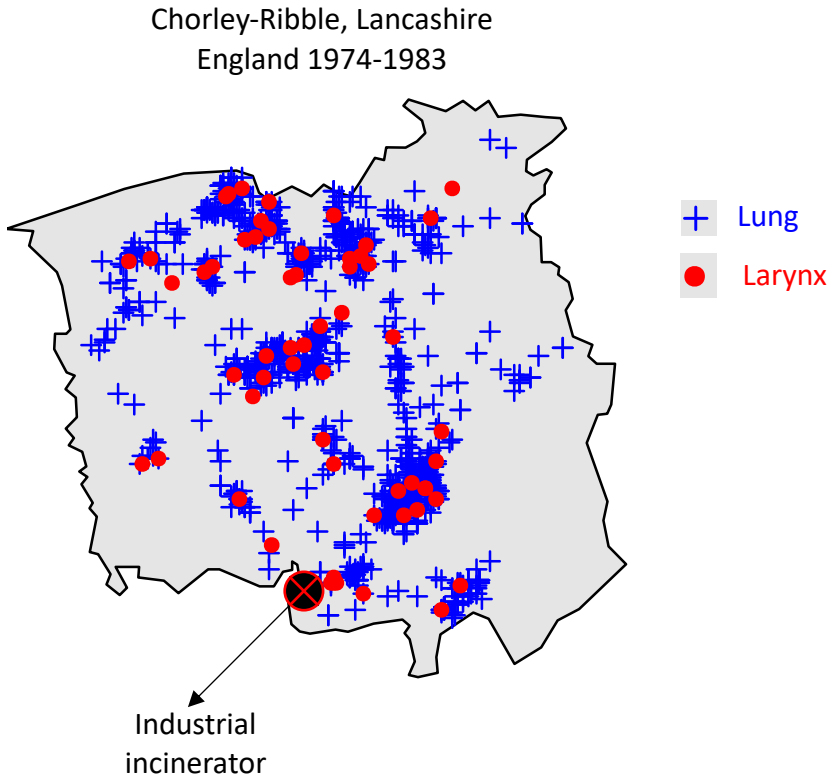
Does the incinerator influence larynx cancer incidence?

- Lung cases as a surrogate for susceptible population
- Model as two inhomogeneous Poisson point processes
- Intensities: $\lambda^{\text{Larynx}}(\cdot)$ and $\lambda^{\text{Lung}}(\cdot)$

If there is an effect:

1. The ratio $\lambda^{\text{Larynx}}(\cdot)/\lambda^{\text{Lung}}(\cdot)$ would be non-constant
2. It would depend on the distance to the incinerator

Focus of our work is on testing #1



Peter J. Diggle. *A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point*, JRSSA, 1990

Goal

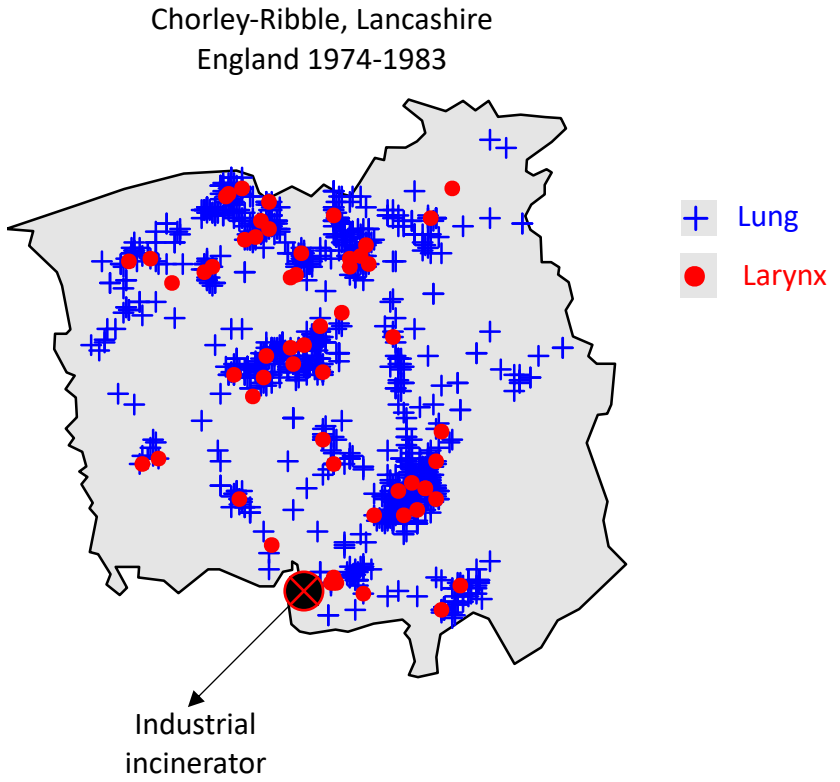
For point processes P and Q, test null hypothesis:

$$\frac{\lambda^P(\cdot)}{\lambda^Q(\cdot)} = \text{const}$$

- Do the intensities of P and Q have the same functional form?

Why not test for $\lambda^P = \lambda^Q$?

- Conflates location pattern with the total number of points in the pattern
- Comparing raw frequency histograms vs. normalized histograms
- For the cancer example:
  - Obviously, fewer red dots: easily reject $\lambda^P = \lambda^Q$
  - Are the location patterns different? Is $\lambda^P/\lambda^Q = \text{const}$?



Two-sample problem

Define location density of events:

$$p(\cdot) = \frac{\lambda^P(\cdot)}{\int \lambda^P(\mathbf{x})d\mathbf{x}} \qquad q(\cdot) = \frac{\lambda^Q(\cdot)}{\int \lambda^Q(\mathbf{x})d\mathbf{x}}$$

Null hypothesis

$$\frac{\lambda^P(\cdot)}{\lambda^Q(\cdot)} = \text{const} \quad \Longleftrightarrow \quad p(\cdot) = q(\cdot)$$

- Nuisance parameter is gone!
- Two-sample problem:
  - Are the two samples drawn from the same distribution?

Existing Approaches

Kelsall & Diggle, 1995a, 1995b:

- Kernel density estimate of the logarithm of intensity ratio

Zhang & Zhuang, 2017:

- Kolmogorov-Smirnov like comparison of masses for a collection of pre-specified regions

$$\sup_{A \in \mathscr{S}} \left| \frac{\mathcal{N}(A)}{\mathcal{N}(\mathcal{S})} - \frac{\mathcal{N}'(A)}{\mathcal{N}'(\mathcal{S})} \right|$$

Fuentes-Santos & González-Manteiga & Mateu, 2017:

- L_2-distance between the kernel density estimates of p and q

Use general two-sample methodology:

- Maximum Mean Discrepancy, Wasserstein distance, Energy distance, etc.

Main issues

- Resampling needed to compute p-values, except Zhang & Zhuang
  - Multiple testing at industrial scale, granularity of p-values
  - Real-time systems, visualization
- p-value crisis: no alternative measures such as Bayes Factors available for these methods
- No replicated pattern comparison: e.g. sets of crime patterns on Mondays vs. Wednesdays

Inspiration

Maximum Mean Discrepancy (MMD)

- Notion of dissimilarity between probability distributions
- For a kernel, such as $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}$

$$\mathrm{MMD}^2(p, q) = \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim p}[k(\mathbf{x}, \mathbf{x}')] - 2\mathbb{E}_{\mathbf{x}\sim p,\mathbf{y}\sim q}[k(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{y},\mathbf{y}'\sim q}[k(\mathbf{y}, \mathbf{y}')]$$

- $\mathrm{MMD}^2(p, q) = 0$ if and only if $p = q$

Kernel Mean Embedding

- There exist embeddings μ_p and μ_q of distributions p and q: $\mathrm{MMD}^2(p, q) = \|\mu^p - \mu^q\|^2$
- **Problem:**
  - infinite-dimensional
  - implicit

Proposed Approach

Approximate Kernel Mean Embedding (aKME)

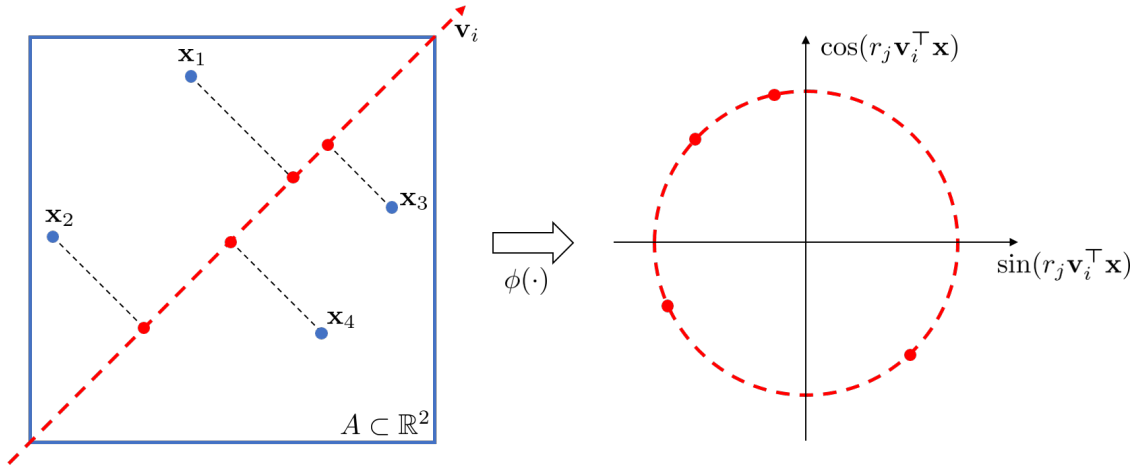
$$\mathrm{MMD}^2(p, q) \approx \|\mathrm{aKME}(P) - \mathrm{aKME}(Q)\|^2$$

- Finite dimensional, explicit formulas, interpretable
- Related to Random Fourier Features (RFF) of Rahimi & Recht 2007
- Custom tailored to 2-dim setting, gives better accuracy than RFF for the same number of dimensions
- Consistency is inherited from MMD
- Not limited to testing based on Euclidean distance (≈MMD^2)
  - easy p-values!

Approach: Step 1

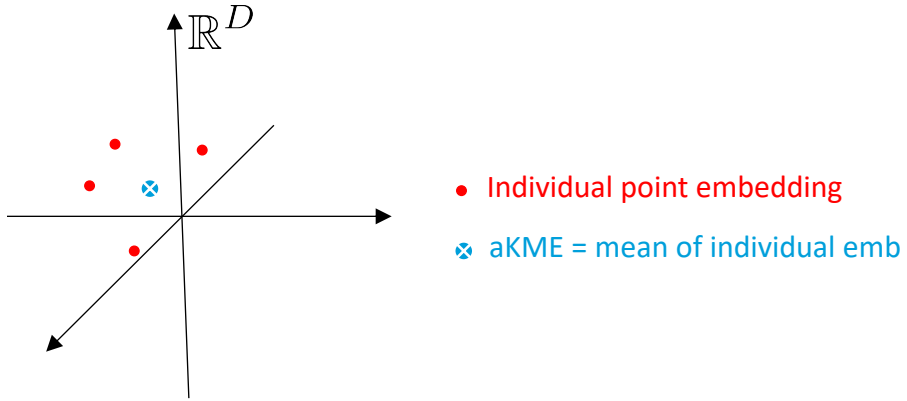
Approximate Kernel Mean Embedding (aKME)

- Pick a line, project all points in the pattern onto the line
- Pick a radius, wrap the line onto the circle of that radius
- Compute sin/cos values and take means
- Result: two number “fingerprint” of the point pattern
  - in given direction
  - at given scale (~circle radius)



Rinse & Repeat:

- Repeat for several lines and radii
- Obtain aKME of the point pattern
- D = 2 x #lines x #radii – dimensional embedding



Approach: Step 2

Test:

$$\frac{\lambda^P(\cdot)}{\lambda^Q(\cdot)} = \text{const} \quad \Longleftrightarrow \quad \text{aKME}(P) = \text{aKME}(Q)$$

- Each dimension of aKME is a mean, so test for equality of means
- Hotelling's T^2 or newer tests such as Chen & Qin do not work
  - Estimating covariance matrix is unstable
  - Nonlinear functional relationships between coordinates of aKME: $\sin^2 + \cos^2 = 1$ and higher order

Approach: Step 2

Our approach

- Apply two-sample **t-tests** on each coordinate, get p-values p_1, p_2, …., p_D
- Combine all the p-values
  - Positive dependencies between tests: cannot use Fisher's combo, Stouffer's Z, …
  - Use the Harmonic Combination or Cauchy Combination

$$p^H = H\left(\frac{D}{\frac{1}{p_1}+\frac{1}{p_2}+\cdots+\frac{1}{p_D}}\right)$$

Wilson, PNAS 2019

$$p^C = \frac{1}{\pi}\cot^{-1}\left(\frac{\cot\pi p_1+\cot\pi p_2+\cdots+\cot\pi p_D}{D}\right)$$

Liu & Xie, JASA 2020

Mean Bayes Factor $\overline{\mathrm{BF}}_{10}$

- Can compute Bayes factors for each t-test using default priors
- Combine resulting BFs via arithmetic mean – optimal according to Vovk & Wang, Arxiv 2019

Simulations: Poisson processes

Inhomogeneus Poisson processes

- Both combination approaches control the size of test
- More powerful when compared to Zhang & Zhuang, JMVA 2017
- Zhang & Zhuang is the only previous test that doesn't require resampling to compute p-values

					This Paper			Zhang-Zhuang		
	Model	β_1	β_2	λ	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Size	Linear	1	1	100	0.011	0.056	0.109	0.008	0.036	0.064
			1	400	0.016	0.059	0.109	0.009	0.040	0.081
			1	800	0.008	0.054	0.094	0.007	0.046	0.086
		2	2	100	0.008	0.056	0.108	0.006	0.038	0.075
			2	400	0.014	0.056	0.098	0.008	0.040	0.084
			2	800	0.008	0.048	0.108	0.012	0.056	0.098
		3	3	100	0.010	0.052	0.100	0.007	0.034	0.077
			3	400	0.011	0.050	0.098	0.010	0.043	0.090
			3	800	0.016	0.062	0.101	0.007	0.040	0.081
	Sine	1	1	100	0.015	0.062	0.106	0.006	0.032	0.068
			1	400	0.014	0.057	0.102	0.012	0.046	0.087
			1	800	0.010	0.056	0.104	0.006	0.042	0.096
		2	2	100	0.010	0.059	0.104	0.004	0.031	0.067
			2	400	0.008	0.048	0.087	0.009	0.040	0.088
			2	800	0.011	0.052	0.092	0.008	0.040	0.086
		3	3	100	0.010	0.060	0.106	0.008	0.038	0.076
			3	400	0.014	0.056	0.104	0.011	0.044	0.086
			3	800	0.008	0.053	0.088	0.010	0.046	0.088
Power	Linear	1	2	100	0.082	0.218	0.320	0.090	0.226	0.332
			2	400	0.656	0.826	0.886	0.638	0.827	0.890
			2	800	0.976	0.996	0.998	0.956	0.987	0.995
			3	100	0.602	0.780	0.839	0.517	0.754	0.844
			3	400	1.000	1.000	1.000	1.000	1.000	1.000
			3	800	1.000	1.000	1.000	1.000	1.000	1.000
		2	3	100	0.069	0.184	0.266	0.058	0.181	0.268
			3	400	0.527	0.740	0.809	0.488	0.718	0.809
			3	800	0.932	0.976	0.988	0.886	0.961	0.981
	Sine	1	2	100	0.412	0.636	0.734	0.137	0.346	0.477
			2	400	1.000	1.000	1.000	0.958	0.992	0.997
			2	800	1.000	1.000	1.000	1.000	1.000	1.000
			3	100	0.964	0.990	0.994	0.568	0.802	0.888
			3	400	1.000	1.000	1.000	1.000	1.000	1.000
			3	800	1.000	1.000	1.000	1.000	1.000	1.000
		2	3	100	0.088	0.232	0.345	0.014	0.057	0.108
			3	400	0.812	0.923	0.952	0.114	0.302	0.436
			3	800	0.996	1.000	1.000	0.494	0.754	0.857

Simulations: Non-Poisson processes

When Poisson assumption is violated:

- Inhibition/Hardcore: more conservative
- Clustering: anti-conservative

	Model	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Size	Hardcore-1	0.012	0.046	0.090
	Hardcore-2	0.008	0.035	0.073
	Hardcore-3	0.002	0.009	0.020
	Cluster-1	0.056	0.164	0.271
	Cluster-2	0.160	0.372	0.508
	Cluster-3	0.413	0.686	0.780

Use effective sample size:

- Divide the total number of points by per-cluster count
- Brings down the size to nominal level
- Useful when e.g. locations from the same user:
  - Effective sample size = #users

	Model	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Size	Cluster-1	0.011	0.051	0.086
	Cluster-2	0.014	0.048	0.086
	Cluster-3	0.015	0.050	0.078

Simulation: Replicated Pattern Comparison

Poisson assumption is not needed

Technical assumption, that can be checked using domain knowledge

	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Size	0.008	0.045	0.090
Power	0.764	0.900	0.938

Application 1: Cancer Data

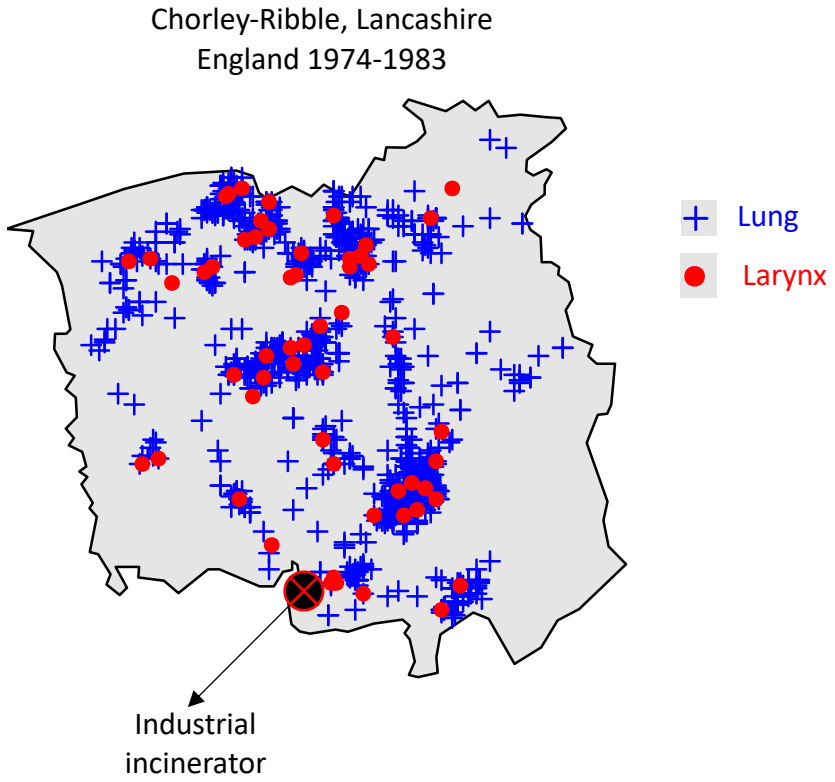
Reanalysis of cancer data

p-value = 0.905

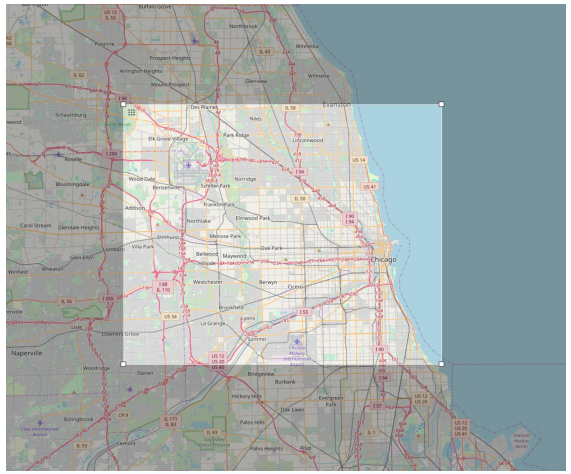
$\overline{\mathrm{BF}}_{10}$ = 0.28

Retain the null

- In agreement with previous re-analyses of this data



Application 2: Chicago Crime



- A year of data, one pattern per **crime type** per **day**
- Replicated pattern comparison
- Different categories of crimes by day of week
- Tue vs Thu is a control, we do not expect to see differences here

Crime Type	Tuesday	Thursday	Saturday	Tue vs Thu		Tue vs Sat	
	#Pts	#Pts	#Pts	p-value	$\overline{\text{BF}}_{10}$	p-value	$\overline{\text{BF}}_{10}$
Theft	144.7	150.3	148.6	0.125	0.838	**0.001**	**29.503**
Battery	88.0	87.2	115.5	0.179	0.682	**0.002**	**12.403**
Deceptive Practice	43.0	42.8	36.2	0.825	0.409	**0.006**	**5.728**
Criminal Damage	50.5	53.6	61.4	0.566	0.502	**0.014**	**2.799**
Robbery	19.9	19.4	22.9	0.099	0.858	0.051	1.231
Narcotics	28.1	27.1	30.1	0.915	0.392	0.613	0.486
Burglary	25.9	25.4	22.8	0.475	0.523	0.657	0.469
Assault	40.3	41.4	37.9	0.987	0.342	0.725	0.435
Other Offense	32.8	31.6	28.1	0.788	0.416	0.968	0.358
Motor Veh Theft	21.1	20.3	22.6	0.997	0.322	0.997	0.322

Rejected @ FDR 0.1

Strong evidence

Substantial evidence

Barely mention!

Summary

Method for point pattern comparison

- Inhomogeneous Poisson point processes
- Works for non-Poisson processes if *effective* sample size can be estimated
- Can do replicated pattern comparison for very general class of processes

Gives a notion of strength of evidence

- Our Mean Bayes Factor can be used to judge evidence against the null
- In line with recent suggestions on reproducibility

Big Data:

- Highly efficient computation
  - aKME can be computed in parallel for each pattern
  - For testing need just basic summaries per aKME dimension: number of points, mean, standard deviation

Future Directions

Production use:

- Can we implement this on Hive/Spark?
  - P-value computation requires t-distribution CDF
  - For combining p-values one can use Cauchy combination which only requires the cotangent function
  - Bayes factor is more complex (numerical integration), maybe use asymptotics?

Visualization: Can we highlight where the point patterns are different?

- Simple approach:
  - Estimate the normalized intensity of point patterns (e.g. kernel smoothing)
  - Visualize the difference
- Hierarchical testing zoom-in:
  - Normalization?
  - How to avoid loss of testing power with smaller samples?

Thank you!

Also check out an e-poster from AT&T Labs

@Thursday Morning session

CatViz:
Visual Exploration of High-Dimensional Categorical Datasets

Eleftherios Koutsofios, Gordon Woodhull, James Klosowski, Raif Rustamov*
Data Science and AI Research, AT&T Labs

*raifrustamov@gmail.com

AT&T DATA SCIENCE & AI RESEARCH

AT&T