Spatial Statistics ~~vs~~ & Machine Learning

Why and how you should consider spatial autocorrelation?

Lyndsay Shand[1]

[1]Department of Statistical Sciences
Sandia National Laboratories

Sandia ML/DL Workshop, July 25-28, 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

DE-NA-0003525. SAND NO. 2022-00000

Perceived Uses/Objectives

Machine Learning

- Focus on ”wide” data: more variables than observations
- Often used to uncover relations across variables to reduce the dimension of the data
- Generally considered predictive models
- Known to be computationally efficient

Statistical Methods

- Focus on ”long” data: more observations than variables
- Require replicates which are often hard to come by - especially in the case of spatial and space-time data
- Generally considered informative models to answer ”how” and ”why” in addition to predictive
- Considered computationally expensive

Spatial Data

Any dataset which can be mapped, that is, has geographical coordinates associated with each observed measurement, is considered spatial data.

Spatial Data

Spatially referenced data can can quickly increase the size of your data making machine learning methods attractive.

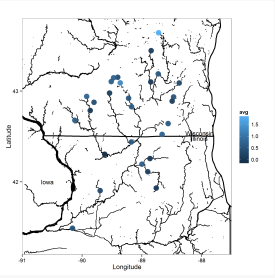
Spatial Data

Spatially referenced data can can quickly increase the size of your data making machine learning methods attractive.

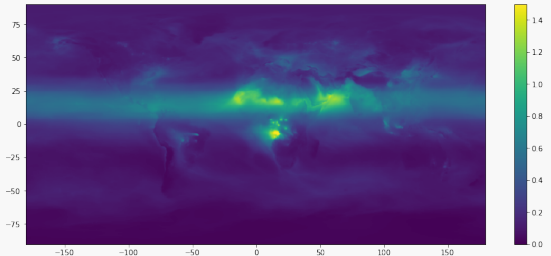
BUT, all spatial data has some degree of spatial autocorrelation which traditional ML methods such as regression, unsupervised clustering, neural networks, ignore.

Recognizing Spatial Data

Geospatial Data.



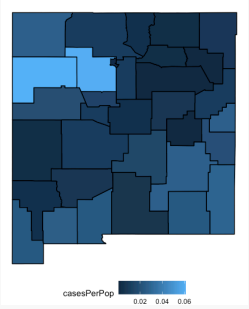
Average river Flow at midwest stations.



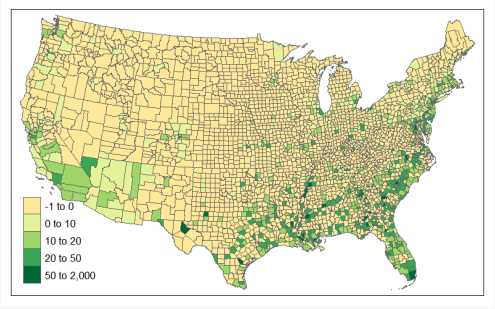
Aerosol Optical Thickness from MERRA-2 Reanalysis `https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2`

Recognizing Spatial Data

Lattice/Areal Data.



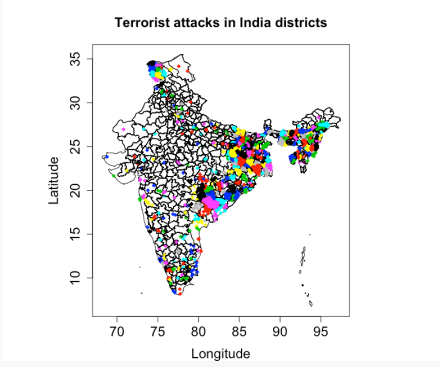
New daily Covid cases by population on October 21, 2020.



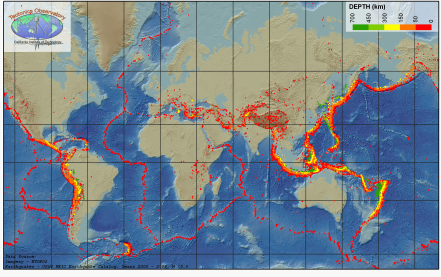
2012 new HIV diagnosis rates in cases per 100,000 across the United States.

Recognizing Spatial Data

Spatial Point Data.



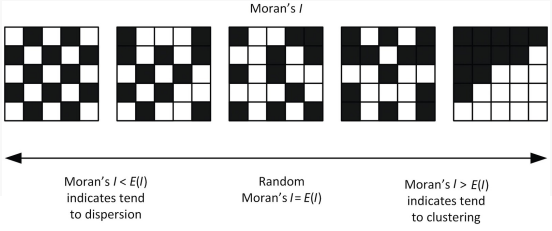
Locations of terrorist attacks in India.
`https://www.start.umd.edu/gtd/`



USGS earthquake catalogue from 2000 to 2008, magnitude of 5.0 M and above.
`https://www.nsf.gov/news/mmg/mmg_disp.jsp?med_id=64691`

Spatial Autocorrelation

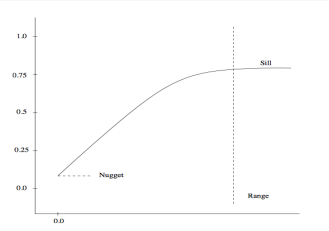
- Spatial autocorrelation refers to similarities across observations due to their physical distance from each other.
- The general assumption is that ”nearest neighbors” have similar characteristics.
- Spatial autocorrelation (similar to temporal autocorrelation) is often overlooked.
- Ignoring spatial autocorrelation, regardless of method, can lead to poor predictive performance and false inference on variable impact.



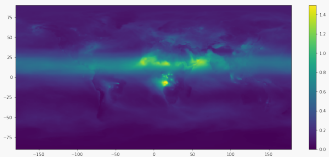
Grekousis, G. (2020)

Spatial Autocorrelation vs. Spatial Trend

- Spatial autocorrelation refers to the correlation between any two locations, typically based on the distance between them.
  - the correlation between locations expressed as a function of distance
  - allows for interpolation at unobserved locations
- Spatial trend refers to the mean trend across a spatial region, usually estimated with spatial replicates
  - can be a function of spatial units, e.g. latitude/longitude



Sample Variogram: C(h=0)-C(h) as a function of spatial distance h.



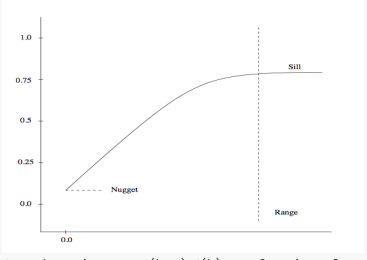
Monthly average Aerosol Optical Thickness

When should we care about spatial autocorrelation?

- climate data
- environmental data
- disease outbreak
- material surface behavior
- road maps
- remote sensing data
- animal movements
- ocean dynamics
- extreme event patterns
- geological patterns
- mining, oil drilling
- agricultural applications

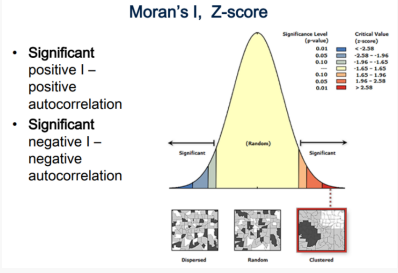
Testing for Spatial Autocorrelation

Variogram for **geospatial data**



Sample Variogram: C(h=0)-C(h) as a function of spatial distance h.

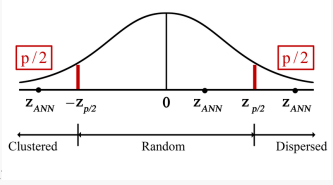
Moran's I or Geary's C for **areal data**



- **Significant** positive I – positive autocorrelation
- **Significant** negative I – negative autocorrelation

https://gis.stackexchange.com/questions/420265/interpreta of-global-moran-i-values-using-pysal

Hypothesis test against Complete Spatial Randomness (CSR) for **point data**



Use Cases

- Characterizing distance-decay properties: How large is the seismic signal of an earthquake as you move away from the source?
- Identifying spatial hot spots: Is there a clear spatial pattern/clustering of covid cases?
- Spatial interpolation: We cannot exhaustively sample the earth. How can we use existing measurements to estimate at unobserved (maybe hard to measure) locations?

How to account for spatial autocorrelation?

Spatial autocorrelation is typically captured in a model's variance structure, e.g. when the i.i.d. error assumption is violated.

For example, in a simple regression framework

$$Y(t,s) = X(t,s)\beta + \phi(s) + \epsilon(t,s), \quad \epsilon(t,s) \overset{iid}{\sim} N(0,\sigma^2)$$

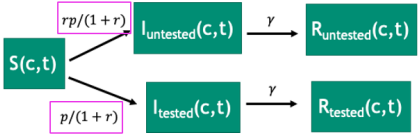
where $\phi(s) \sim MVN(0,\Sigma)$ is a spatial random effect with spatial covariance Σ.

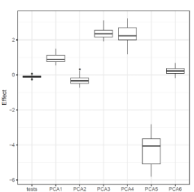
A New Mexico Covid-19 Model

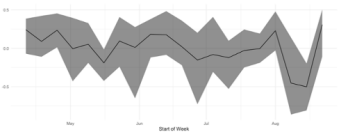
Level II - Linear mixed model to describe the change in **log odds** of incidence rate $p(c,t)$

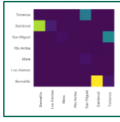
$$logit(p(c,t)) - logit(M(c,t) = Z(c,t), \quad M(c,t) = I(c,t)S(c,t)/N(c)^2$$

$$Z(c,t) = \sum_{k}^{K} x_k(c,t)\alpha_k + \theta_t + \phi_c$$





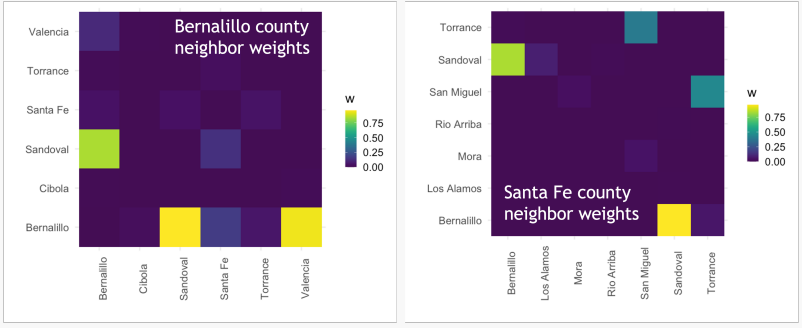




Spatial random effect to capture differences and relationships between regions not already captured by x_k

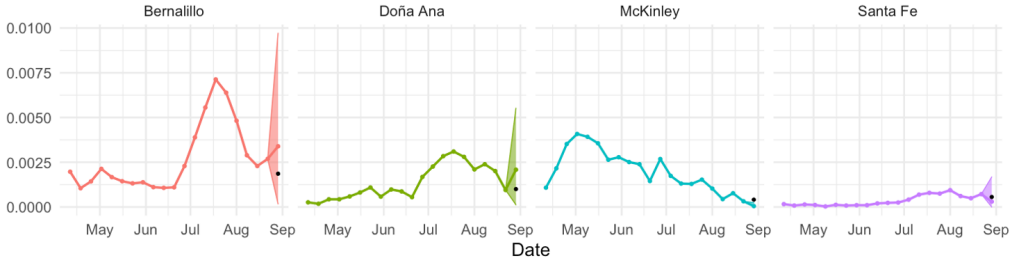
A New Mexico Covid-19 Model

To account for spatial dependency in an intuitive way, we defined the level of influence one county's case numbers has on another as a function of the number of the number of US+State highways that connect any two counties weighted by the level of commuter traffic.

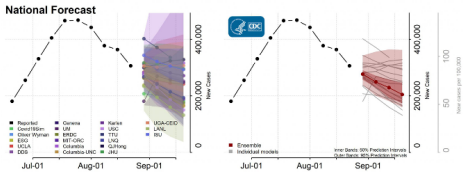


A New Mexico Covid-19 Model

One-week ahead new (tested) cases/100,000 with uncertainty compared to observed (tested) incidence(black dot)



Region specific uncertainty rather than global uncertainty can be more informative for decision makers.



Spatially-Varying AR(1)

Shand et al. 2018 specifies an AR(1) to model HIV spread across the U.S.

$$y_{i,t} = X_{i,t-1}^T \beta + \rho_i Z_{i,t-1} X_{i,t-2}^T \beta, \quad i = 1, \ldots, n \text{ spatial locations}$$

where autocorrelation parameter ρ_i was found to be spatially correlated

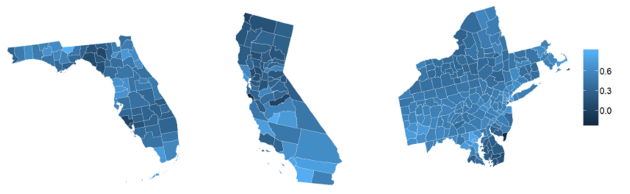
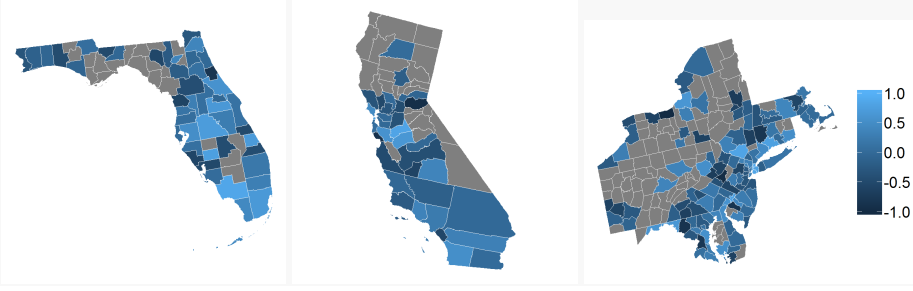


Figure 3.2: Spatial maps of posterior means of ρ_i from model (3.3.2) for Florida, California and New England.

Spatially-Varying AR(1)

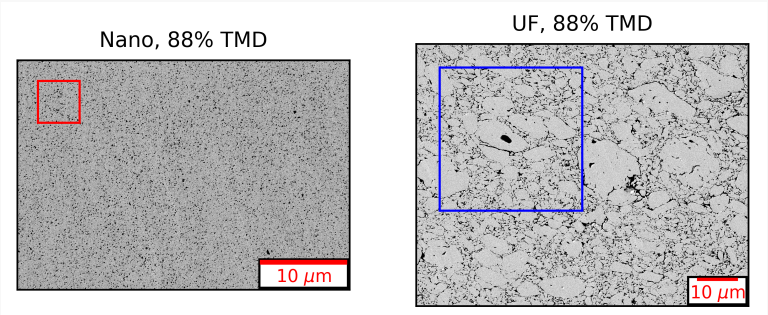
Test Statistics and p-values for for Moran's I and Geary's C testing the null hypothesis of no spatial correlation and the alternative hypothesis of positive spatial correlation.

	Florida		California		New England	
	Statistic	p-value	Statistic	p-value	Statistic	p-value
Moran's I	0.0343	0.3216	0.1072	0.1389	0.2598	0.0003
Geary's C	0.9610	0.3758	0.8098	0.0665	0.7384	0.0005



Spatial maps of independent ρ estimates.

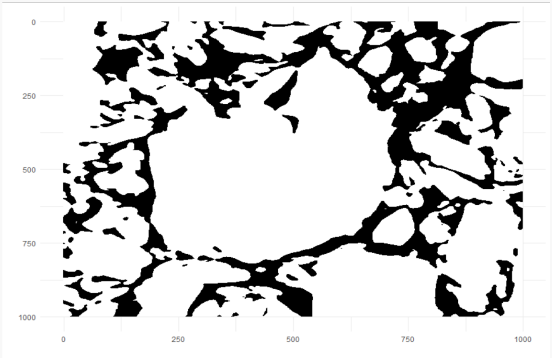
Microstructure Reconstruction

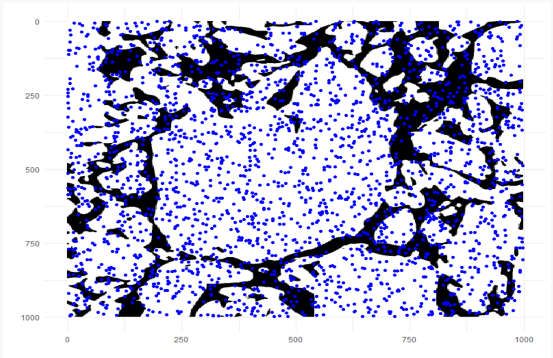


Triaminotrinitrobenzene (TATB) w/ “Nano” (left) and ultra-fine (right) grain size and material density of $\sim 88\%$

Objective: Can we characterize the underlying microstructure well enough to reconstruct an image with similar structural properties of interest?

Microstructure Reconstruction





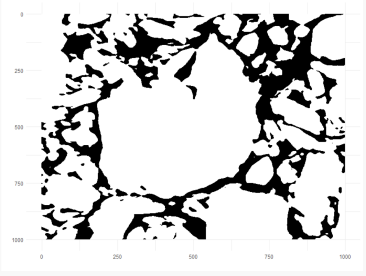
To ease computation:

1. Apply smoothing filter to make 500 x 500 pixel image
2. Sample 1% of points $\rightarrow$ $m = 2,500$ points used to estimate $\boldsymbol{\theta}$ and $\boldsymbol{z}$

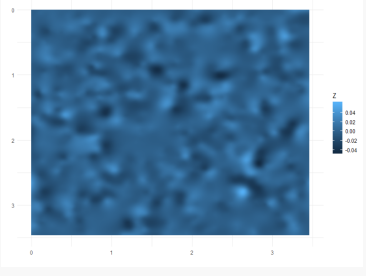
Microstructure Reconstruction

Fitting a Gaussian Process with the nonstationary spatial covariance of Shand and Li (2017) to the sub-selected points, we can reconstruct the original microstructure indicating we are capturing key spatial characteristics.

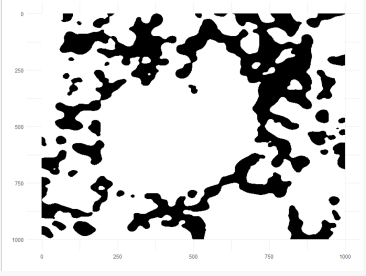
Original



Latent z



Regeneration



Spatial Autocorrelation in Multivariate Climate Variables

Consider the Echo State Network of McDermott and Wikle (2018)

$$\begin{aligned}
\text{Data Stage: } & \boldsymbol{Z}_t \approx \boldsymbol{\Phi}\boldsymbol{\alpha}_t \\
\text{Output Stage: } & \boldsymbol{\alpha}_t = V_1 h_t + V_2 h_t^2 + \eta_t,\ \eta_t \sim N(0, \sigma_\eta^2 I) \\
\text{Hidden Stage: } & h_t = g_h\left(\frac{v}{|\lambda_w|} W h_{t-1} + U\tilde{x}\right)
\end{aligned} \tag{1}$$

where output and hidden stages represent a recurrent neural network and data stage is a dimension reduction technique using spatial EOFs $\boldsymbol{\Phi}$.

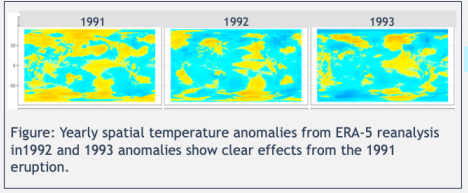
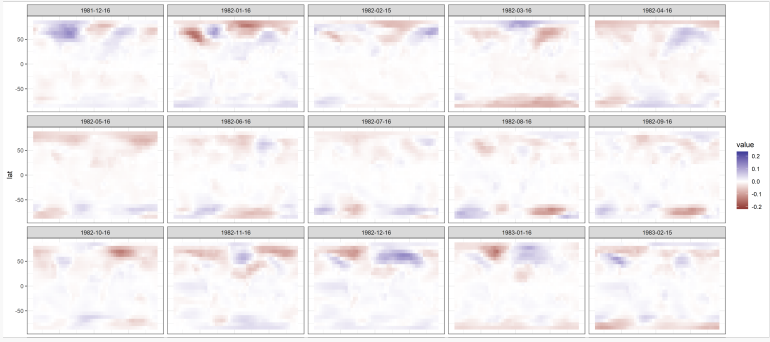


Figure: Yearly spatial temperature anomalies from ERA-5 reanalysis in1992 and 1993 anomalies show clear effects from the 1991 eruption.

Spatial Autocorrelation in Multivariate Climate Variables

Examining $Z_t - \hat{Z}_t$, from fitting the ESN to monthly temperature data, we see that spatial trends and correlations still remain unaccounted for.



Takeaways

- Consider the presence of spatial autocorrelation in your data.
- Is this an important feature of your data?
- How can spatial correlation be accounted for in your modeling approach?
- How can spatial correlation be used to develop more informative/predictive machine learning methods for spatial data?

Not being a machine learning expert myself, I am always eager to learn how ML can incorporate elements from spatial statistics. Please reach out to discuss!

References

Grekousis, G. 2020. Spatial Autocorrelation. In Spatial Analysis Methods and Practice: Describe – Explore – Explain through GIS (pp. 207-274). Cambridge: Cambridge University Press.

Patel, L., Shand L, Tucker, J. D.,Huerta, J. G. 2021. Spatio-temporal extreme event modeling of terror insurgencies. arXiv:2110.08363

Shand L, Li B, Park T, Albarracín D. 2018. Spatially varying auto-regressive models for prediction of new human immunodeficiency virus diagnoses. *J R Stat Soc Series B Stat Methodol.* 67(4):1003-1022.

Shand L, Li B. Modeling nonstationarity in space and time. 2017. *Biometrics.* 73(3):759-768.

Tucker, J. D. Shand, L., Lewis, J. R. 2019 Handling missing data in self-exciting point process models, Spatial Statistics, 29, 160-176.