GEODETECTOR:

CREATING RANDOMNESS AND WORKING WITH HETEROGENEITY IN BIG DATA OR SURVEY DATA

Jinfeng WANG

wangjf@Lreis.ac.cn
State Key Laboratory of Resource & Environmental Information System
Institute of Geographical Science & Nature Resources Research
Chinese Academy of Sciences
Beijing 100101, China

www.geodetector.cn

2022-11-22

Statistics & Spatial Data

Spatial Stratified Heterogeneity (SSH)

Causation of/by SSH

9 Stories

Discussion

Conclusion

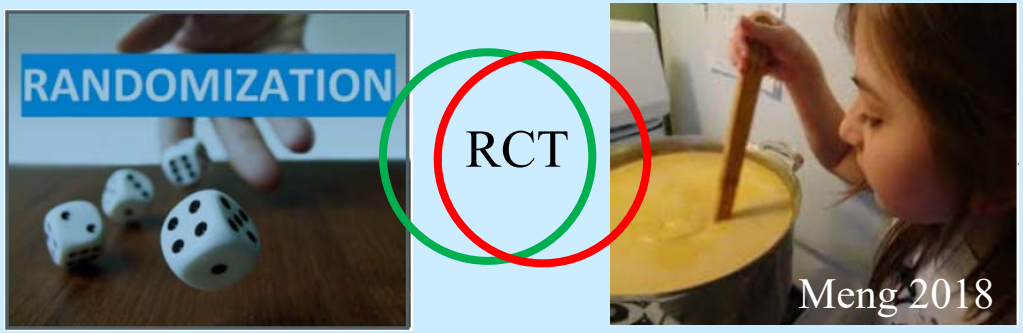
Science has been based on two great achievements: formal logical system and the causal relationships by systematic experiment.

-- 1953, Einstein Archive 61-381

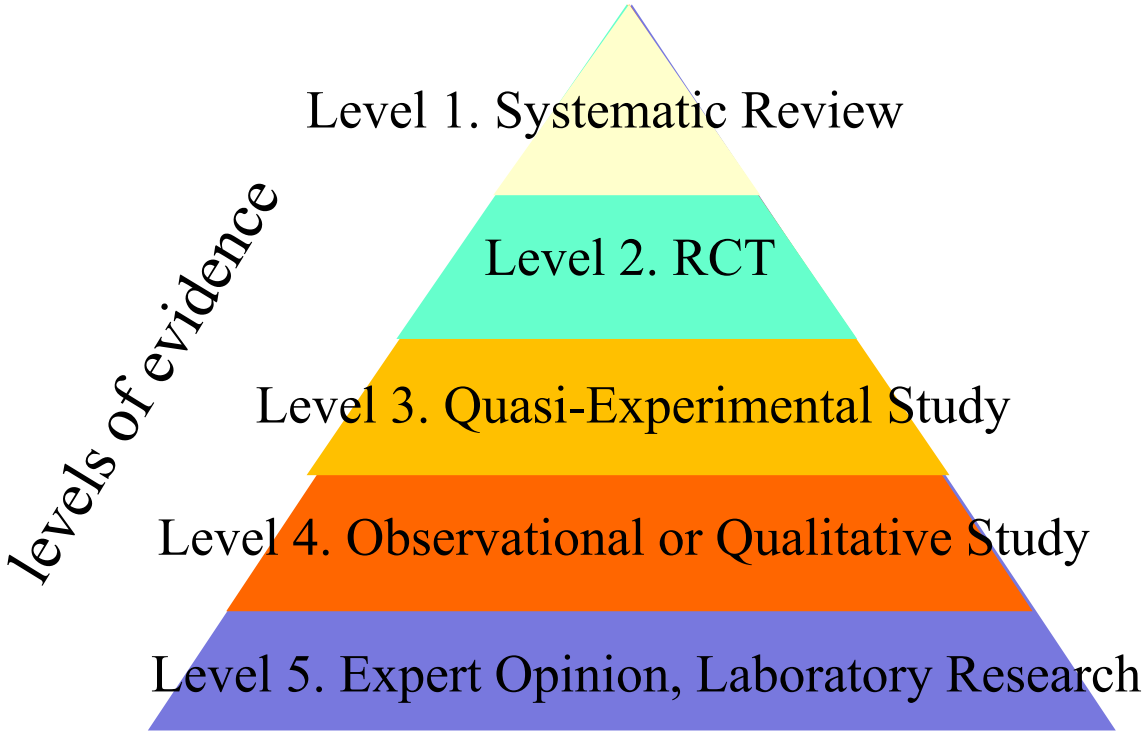
Statistical Experiment

Sampling: **Random & Repeat**

Population: **Identical Distribution**



Medical Data



Sampling:
Random & Repeat

Population:
Identical Distribution

Statistical Experiment

RCT

Neglect the data features

Removed by assumption

Bias; Confounded; Misleading CI

80% of Data Has Spatial Component

(1987 Williams)

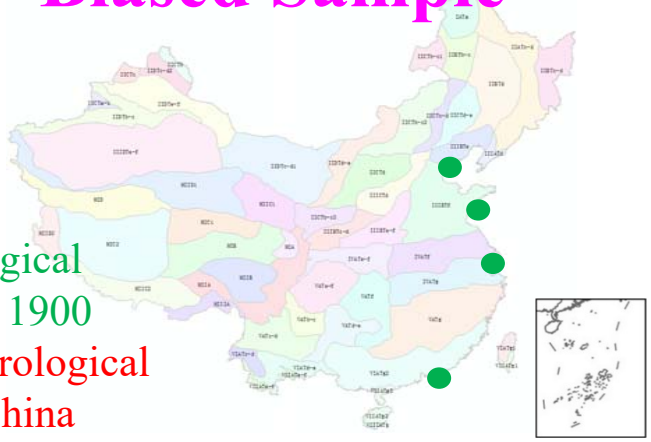
Meteorological stations in 1900 and meteorological zones in China

Sampling:
Neither Random Nor Repeated

Population:
Spatial Stratified Heterogeneity

Wang JF, Zhang TL, Fu BJ. 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 7: 250-256.

Biased Sample



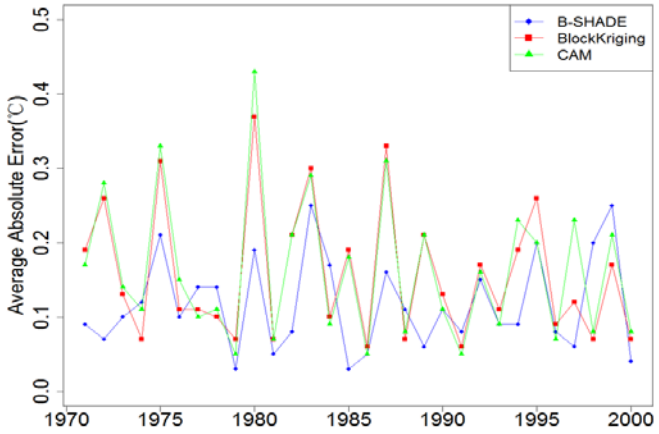
Meteorological stations in 1900 and meteorological zones in China

Confounded

	Treatment A			Treatment B		
	Works	Total	W/T	Works	Total	W/T
Small Stone	81	87	93%	234	270	87%
Large Stone	192	263	73%	55	80	69%
Overall	**273**	**350**	**78%**	**289**	**350**	**83%**

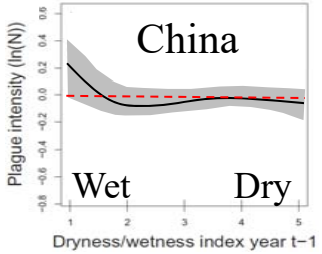
Charig CR, et al. 1986. *Br Med J (Clin Res Ed)* 292 (6524): 879–882.

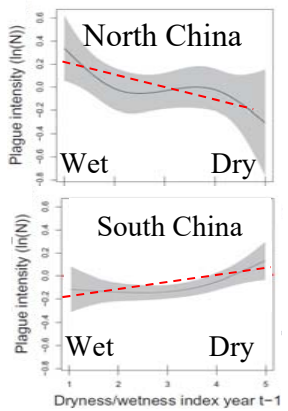
Biased Statistic



Wang JF, Xu CD, Hu MG, *et al*. 2014. *JGR* 119(1): 1-9.

Misleading CI





Xu L, *et al*. 2011. *PNAS*

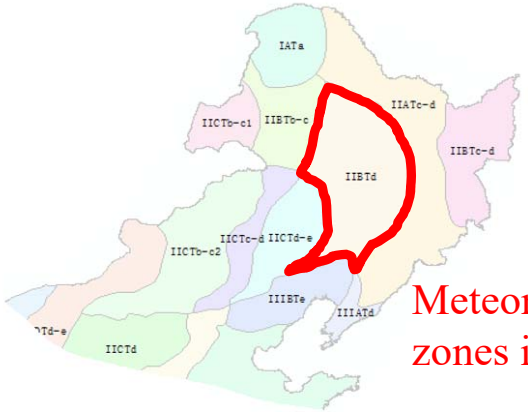
are sourced from:

Spatial Stratified Heterogeneity

Wang JF, Zhang TL, Fu BJ. 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 7: 250-256.

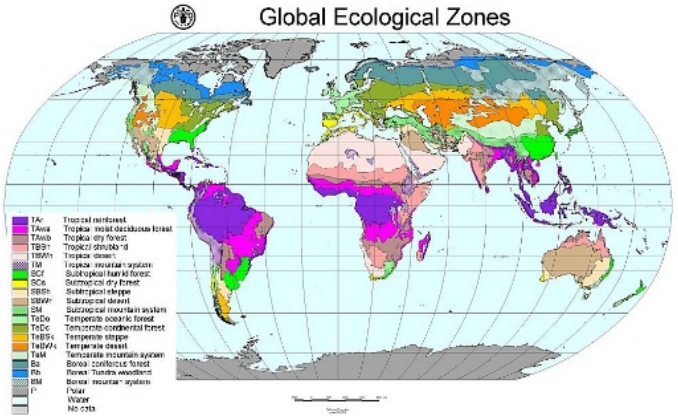
Spatial Stratified Heterogeneity offers benefits:

Stratum: Random

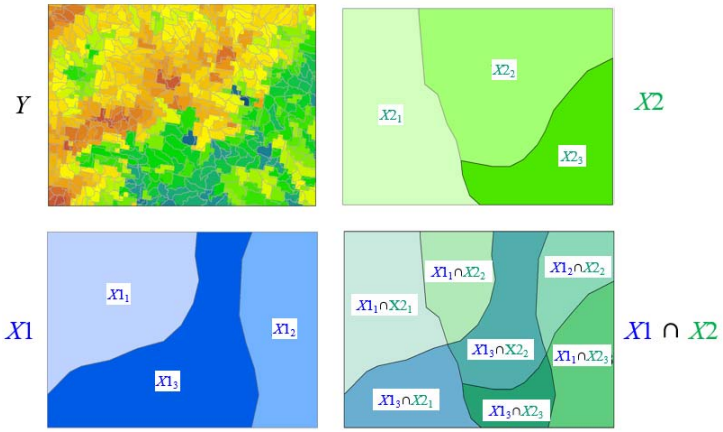


Meteorological zones in China

Strata: Spatial Pattern



Overlay: General Interact



btw Strata: Nonlinear Cause

L Holdridge (1907-1999)

Wang JF, Zhang TL, Fu BJ. 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 7: 250-256.

Statistics & Spatial Data

Spatial Stratified Heterogeneity (SSH)

Causation of/by SSH

9 Stories

Discussion

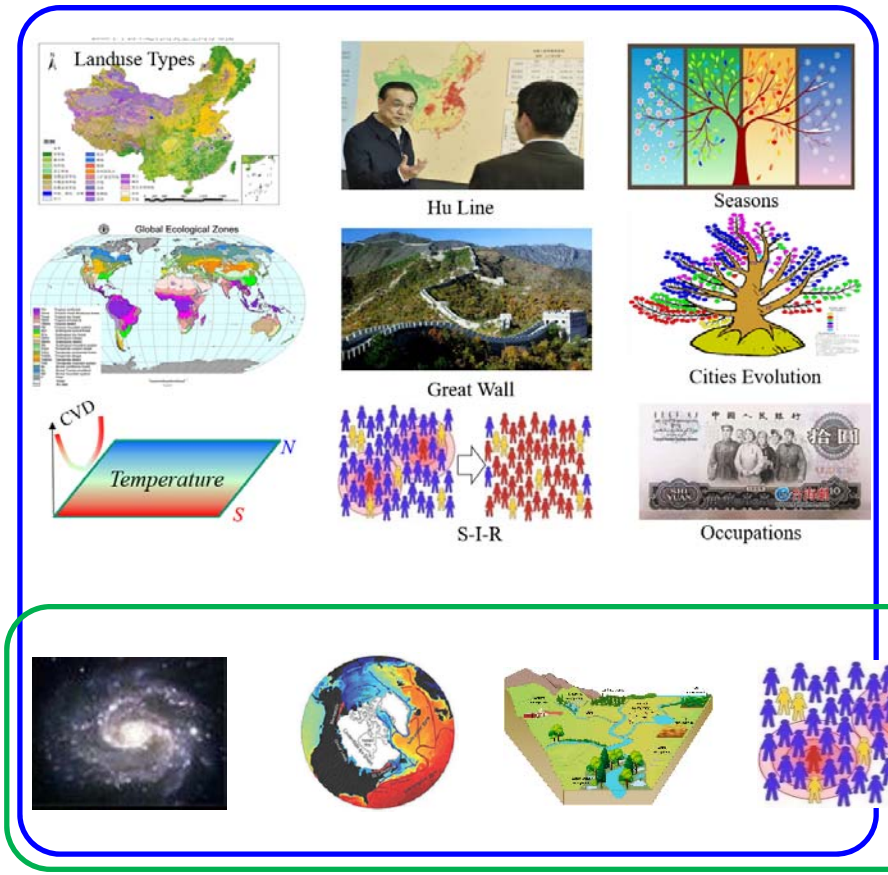
Conclusion

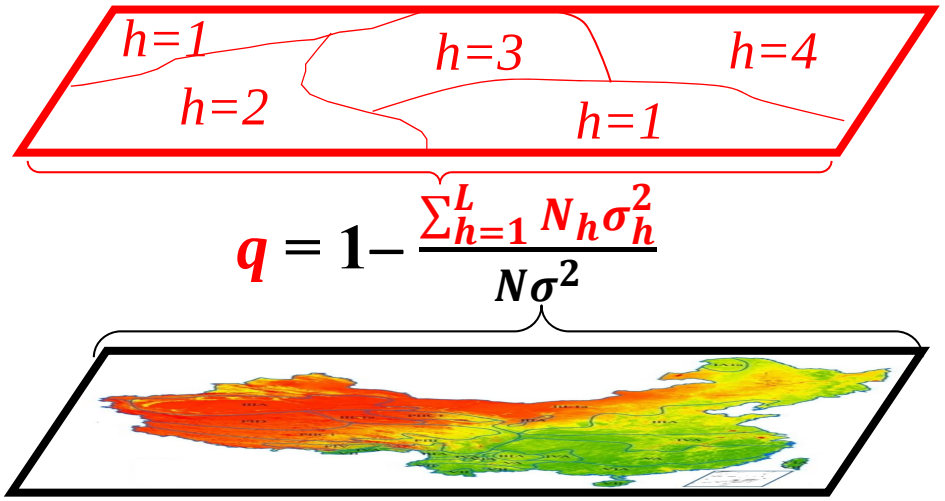
Find an overarching theme or concept with which to structure my heterogeneous materials.

-- Tuan Y-F. Space and Place: The Perspective of Experience
University of Minnesota Press (2001), p.v.

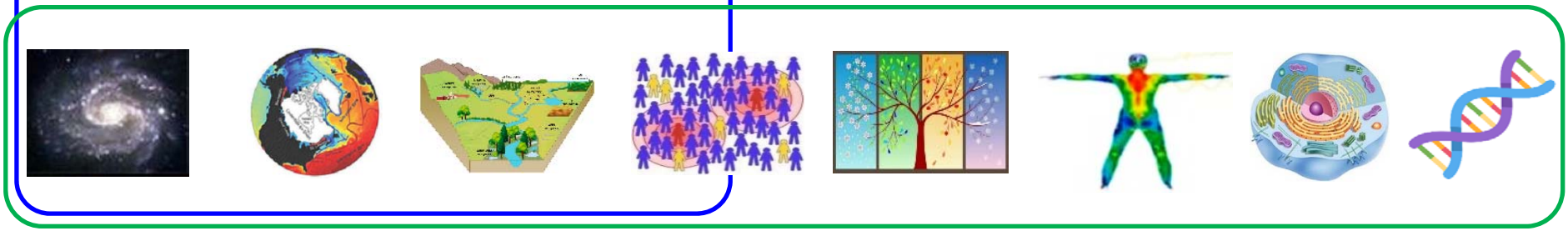
Spatial Stratified Heterogeneity (SSH)

within strata is more similar than between strata





$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2}$$



$q \in [0, 1]$, Y has $100q\%$ degree of SSH

$= 0$, if Y has no SH

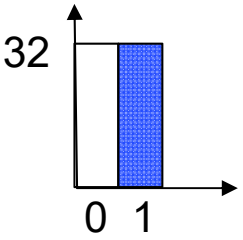
$= 1$, if Y is fully SH

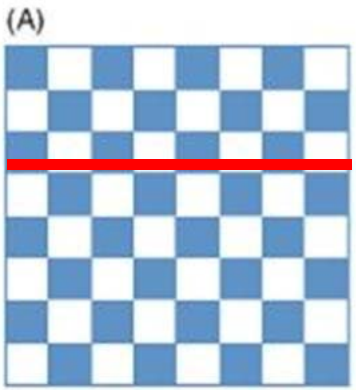
$$\frac{N-L}{L-1}\frac{q}{1-q} \sim F(L-1, N-L; \lambda)$$

Wang JF, Zhang TL, Fu BJ. 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 7: 250-256.

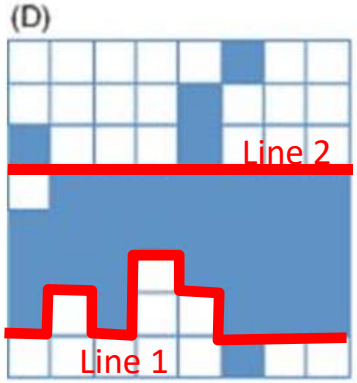
Spatial Stratified Heterogeneity

$N_w = 32$ $\bar{Y} = 0.5$
$N_B = 32$ $\sigma^2 = 0.25$

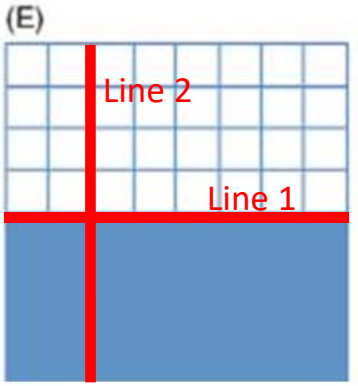




$q = 0$



$q_1 = 0.16^*$
$q_2 = 0.27^{**}$
$q_{1\&2} = 0.67^{***}$



$q_1 = 1^{***}$
$q_2 = 0$

Goodchild M. 1986. CATMOG, GeoBooks, Norwich

Wang JF, Zhang TL, Fu BJ. 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 67: 250-256.

Statistics & Spatial Data

Spatial Stratified Heterogeneity (SSH)

Causation of/by SSH

9 Stories

Discussion

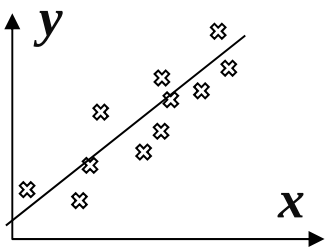
Conclusion

Correlations cry out for explanations.

-- Bell JS. Speakable and Unspeakable in Quantum Mechanics,
Cambridge University Press (1987), p.152.

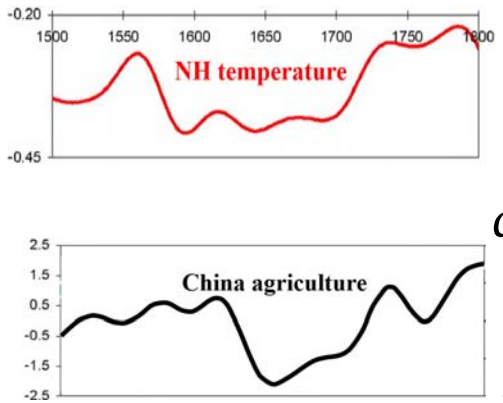
Association + Info → Causation

Correlation btw Two Paints



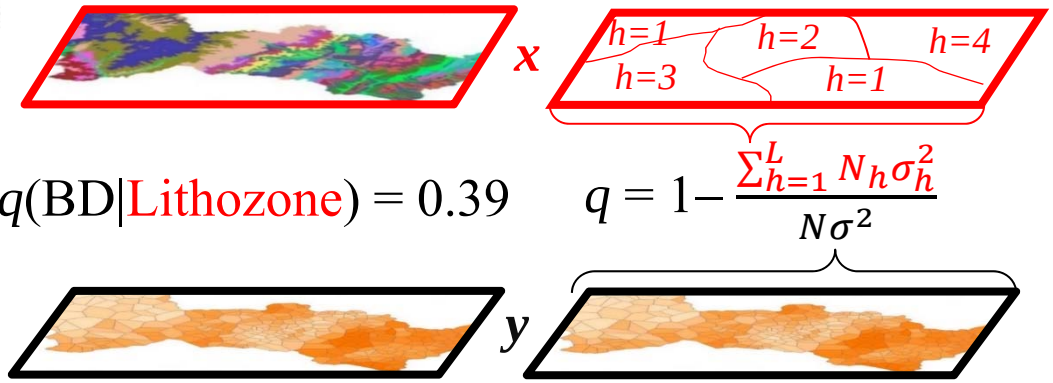
Pearson K. 1895. Notes on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London** 58: 240–242

Coupling btw Two Series



Zhang D, *et al*. 2007. Global climate change, war, and population decline in recent human history. **PNAS** 104(49): 19214–9.

If *X* causes *Y*, the two patterns tend to be coupled



$q(\text{BD}|\text{Lithozone}) = 0.39$

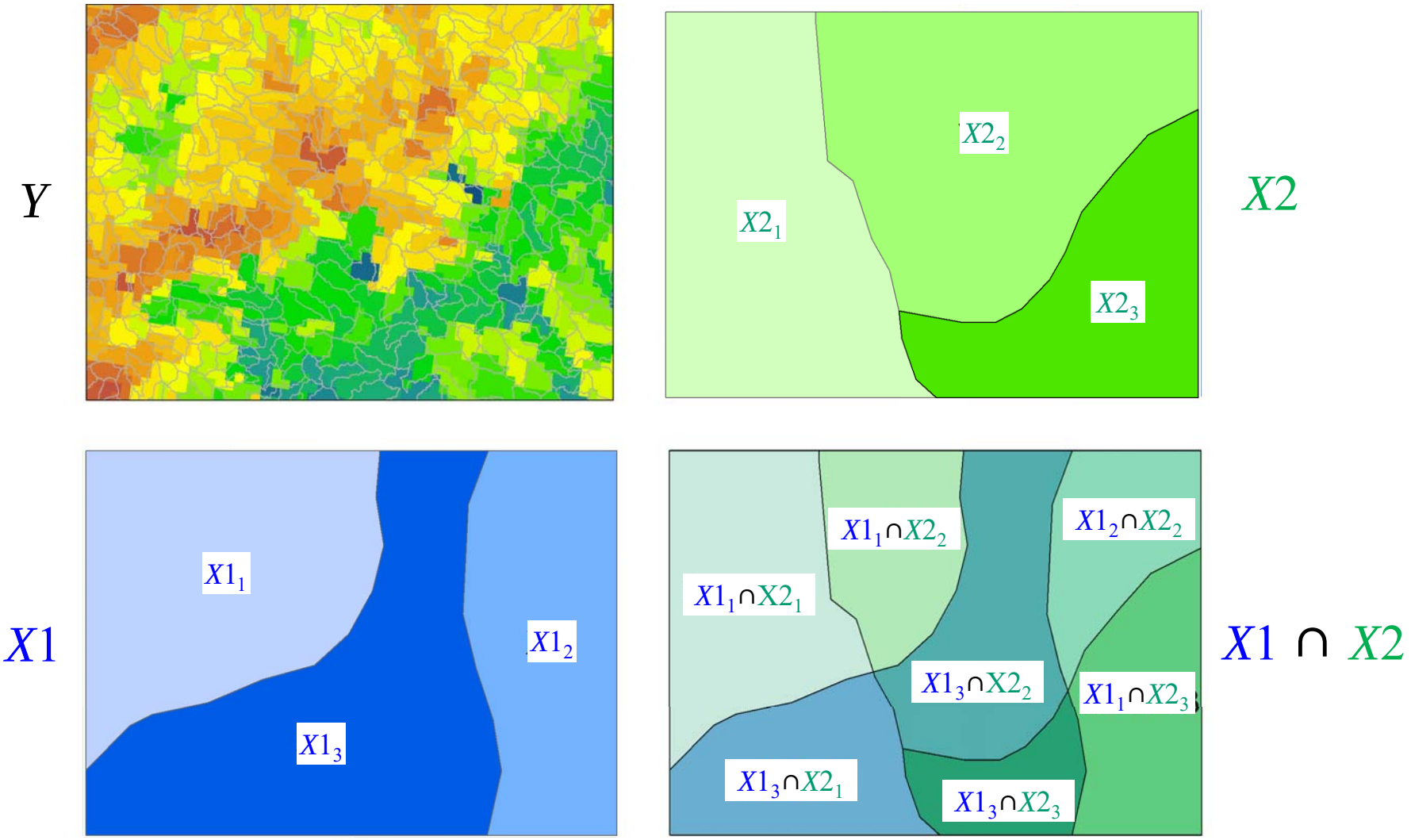
$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2}$$

Wang JF, *et al*. 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. **IJGIS** 24(1): 107-127.

百世修来同船渡　　千世修来共枕眠

Wang JF, Li XH, Christakos G, et al. 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. International Journal of Geographical Information Science 24(1): 107-127.

Interaction: $Y \leftarrow X1 \cap X2$



Wang JF, Li XH, Christakos G, et al 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. International Journal of Geographical Information Science 24(1): 107-127.

Statistics & Spatial Data

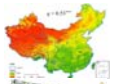
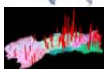
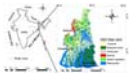
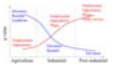
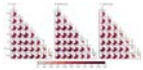
Spatial Stratified Heterogeneity (SSH)

Causation of/by SSH

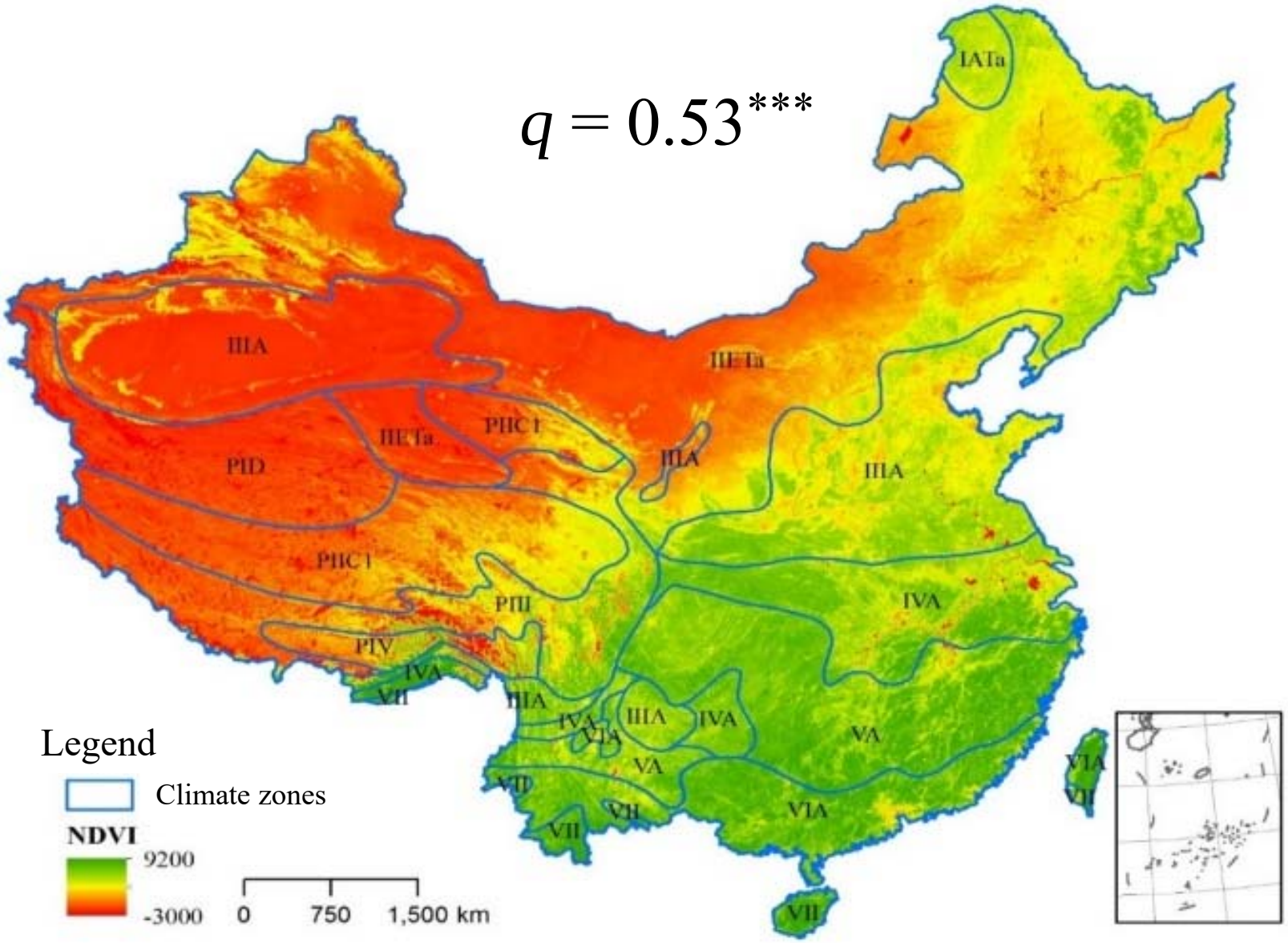
9 Stories

Discussion

Conclusion

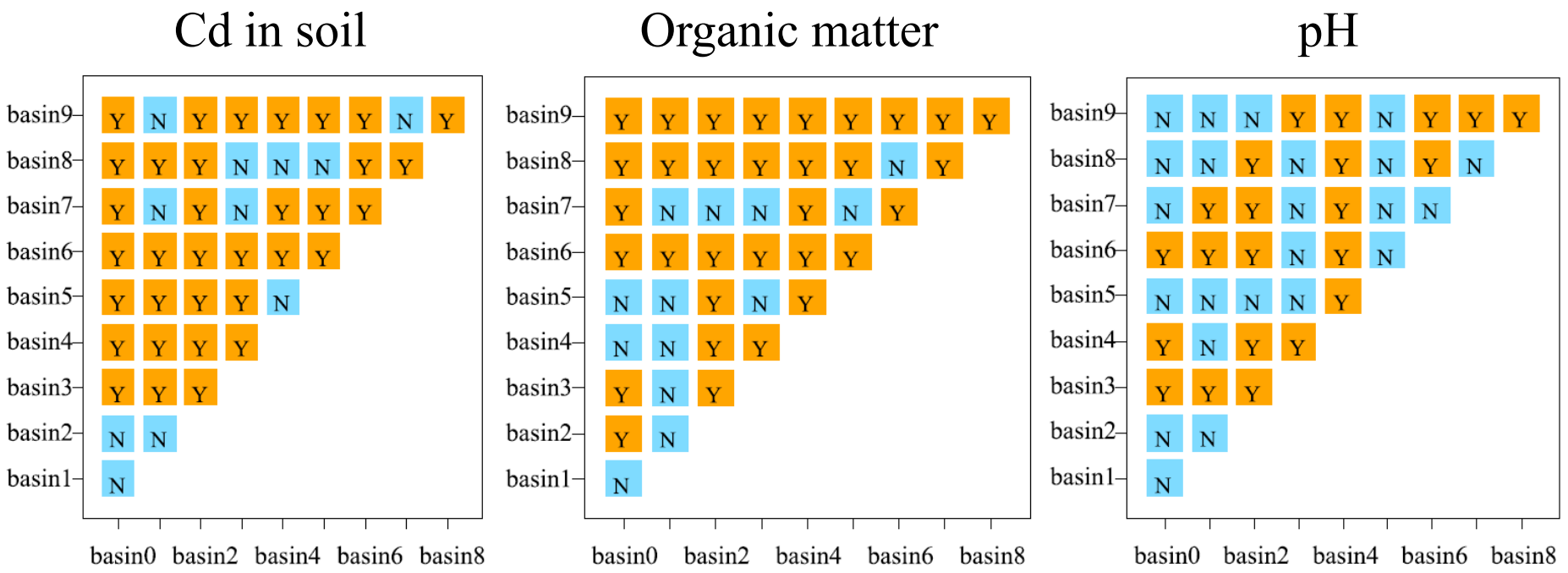
9 Stories		Geodetector *q*
1	Test existence of a spatial pattern (NDVI pattern in China)	 $q(y\|\#)$, # is pattern
2	Find SSH, then modelling in strata (Lung cancer in US)	 $q(y\|\forall\#)$
3	Determinants' spectrum, global (Birth defects in a county)	 $\{x_k\}$: $q_k(y\|\#(x_k))$, $\forall k$
4	Determinants' spectrum, in regions (Land dissection in US)	 $q_{zk}(y\|\#(x_k))$, $\forall k$, $\forall z$
5	Determinants' spectrum, in modules (Three eco-services)	 $q_{mk}(y\|\#(x_k))$, $\forall k$, $\forall m$
6	Determinants' spectrum, evolution (Hu line in 80 years)	 $q_{tk}(y\|\#(x_k))$, $\forall k$, $\forall t$
7	General interaction (Virus and meteorological factors)	 $q(y\|\#(x_1)\cap\#(x_2))$
8	Spatial Goodness of Fit (Global mortality & temperature)	 $q(\#\hat{y}\|\#y)$
9	Geodetector +	
	(1) Kriging (Mapping soil organic carbon)	
	(2) GWR (Spread pattern of COVID-19)	
	(3) InSAR (Ground deformation)	
	(4) Google Earth Engine (Land use changes & drivers)	
	(5) SWAT (Water conservation function)	
	(6) BHM (POI urban vibrancy)	
	(7) …	

Story 1. Spatial Pattern of NDVI



Wang JF, Zhang TL, Fu BJ. 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 7: 250-256.

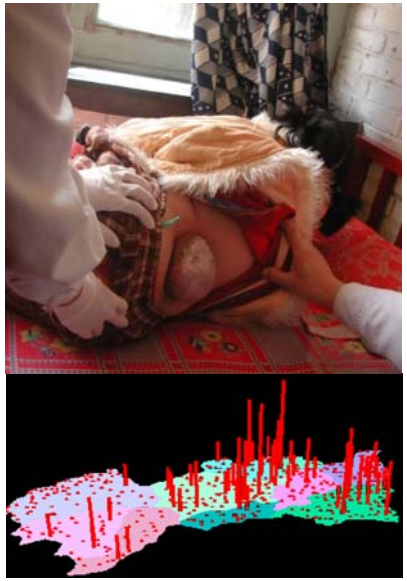
Story 2. Kriging in Strata



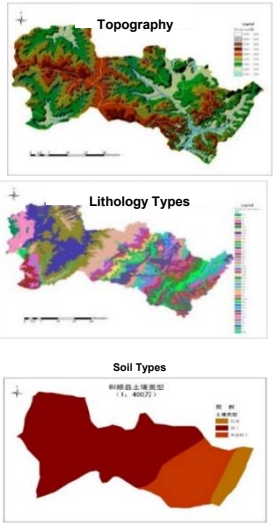
Merge two watersheds if their SSH is insignificant,
so Kriging would be more accurate by avoiding the error in border btw strata

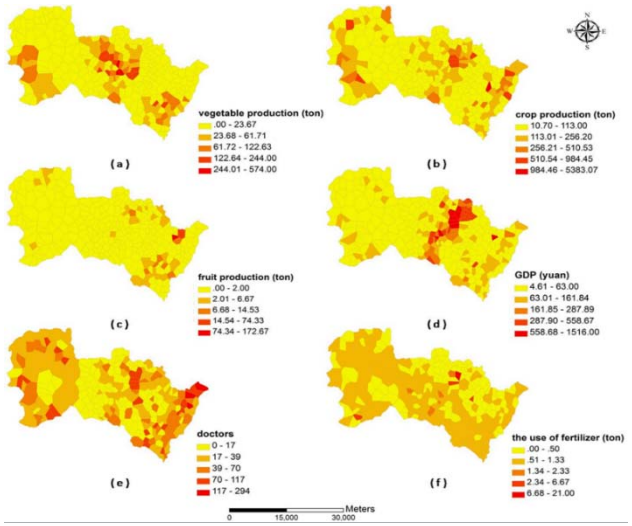
Yang JT, et al. 2022. Chain modeling for the biogeochemical nexus of cadmium in soil–rice–human health system. Environment International 167 (2022) 107424

Story 3. Determinants of Neural Tube Defects



Y ⟵ *Geodetector* ⟶ X

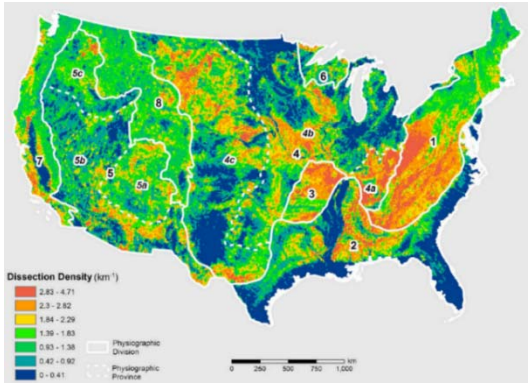




Environment q (watershed 0.47, lithozone 0.39, soil 0.19) controls NTD.
Nutrition q (food 0.18) > Pollution q (fertilizer 0.09) in controlling NTDs.
Interaction: ancient materials released from faults (q = 0.19) then spreading along slopes (q = 0.09) dramatically increase the risk of NTDs (q = 0.86).

Wang JF, Li XH, Christakos G, et al. 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. International Journal of Geographical Information Science 24(1): 107-127.

Story 4. Determinants of Dissection in Regions



Y ← *Geodetector* → X

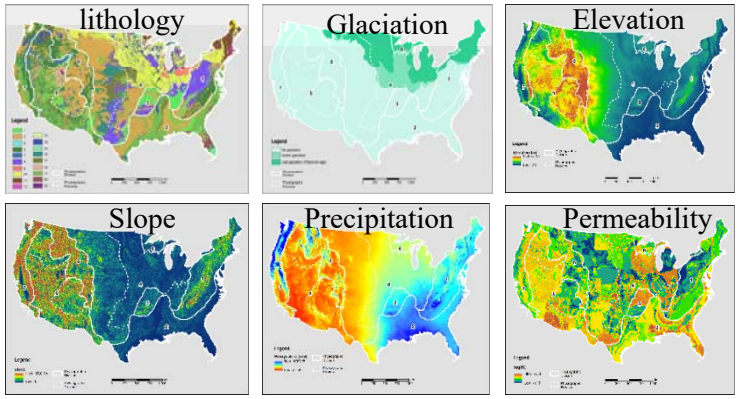


Table 2. Factor or Factor Interaction With Maximum *q* Value

Physiographic Division/Province Name (#)	Dominant Factor	*q*	Dominant Interaction	*q*
Appalachian Highlands (1)	litho	0.49	litho ∩ elev	0.58
Atlantic Plain (2)	planc	0.46	elev ∩ precip[a]	0.62
Interior Highlands (3)	log*k*	0.31	log*k* ∩ precip	0.47
Interior Plains (4)	planc/tanc	0.29	litho ∩ planc/litho ∩ tanc	0.40
Interior Low Plateau (4a)	litho	0.21	log*k* ∩ litho/log*k* ∩ difelev	0.34
Central Lowlands (4b)	glaci	0.36	glaci ∩ planc/glaci ∩ tanc	0.52
Great Plains (4c)	planc	0.24	log*k* ∩ litho/log*k* ∩ tanc	0.34
Intermontane Plateaus (5)	litho	0.31	litho ∩ slp	0.37
Colorado Plateau (5a)	litho	0.13	litho ∩ planc	0.20
Basin and range (5b)	litho	0.35	Litho ∩ slp/litho ∩ tanc	0.41
Columbia Plateau (5c)	litho	0.33	litho ∩ precip	0.46
Laurentian Upland (6)	litho	0.23	litho ∩ tanc	0.32
Pacific Mountain System(7)	elev	0.28	elev ∩ precip[a]	0.43
Rocky Mountain System (8)	litho	0.10	litho ∩ tanc/litho ∩ slp	0.18

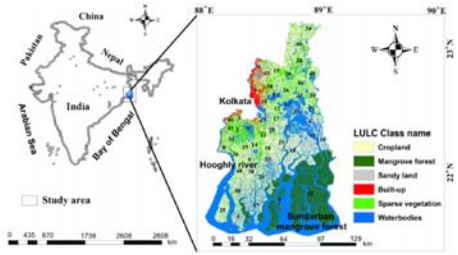
Luo W, Jasiewicz J, Stepinski T, et al. 2016. Spatial association between dissection density and environmental factors over the entire conterminous United States, *Geophys. Res. Lett.* 43: 692–700

Story 5. Determinants of Eco Service in Modules

Individual and joint effects of the driving factors on ESs (a) biophysical, (b) economic, (c) hybrid modules.

	Biop_Factor	Clim_Factor	Dev_Factor	EndStress_Fac	Ldeg_Fac	Soceco_Fac
(a)						
Biop_Factor	0.114					
Clim_Factor	0.774	0.670				
Dev_Factor	0.225	0.713	0.015			
EndStress_Fac	0.680	0.753	0.569	0.509		
Ldeg_Fac	0.649	0.709	0.632	0.581	0.477	
Soceco_Fac	0.620	0.780	0.359	0.684	0.687	0.206
(b)						
Biop_Factor	0.050					
Clim_Factor	0.221	0.051				
Dev_Factor	0.419	0.282	0.202			
EndStress_Fac	0.294	0.300	0.421	0.138		
Ldeg_Fac	0.176	0.094	0.305	0.210	0.019	
Soceco_Fac	0.625	0.460	0.496	0.647	0.528	0.395
(c)						
Biop_Factor	0.122					
Clim_Factor	0.754	0.658				
Dev_Factor	0.225	0.693	0.002			
EndStress_Fac	0.597	0.766	0.490	0.418		
Ldeg_Fac	0.577	0.709	0.589	0.551	0.429	
Soceco_Fac	0.652	0.778	0.437	0.677	0.699	0.240

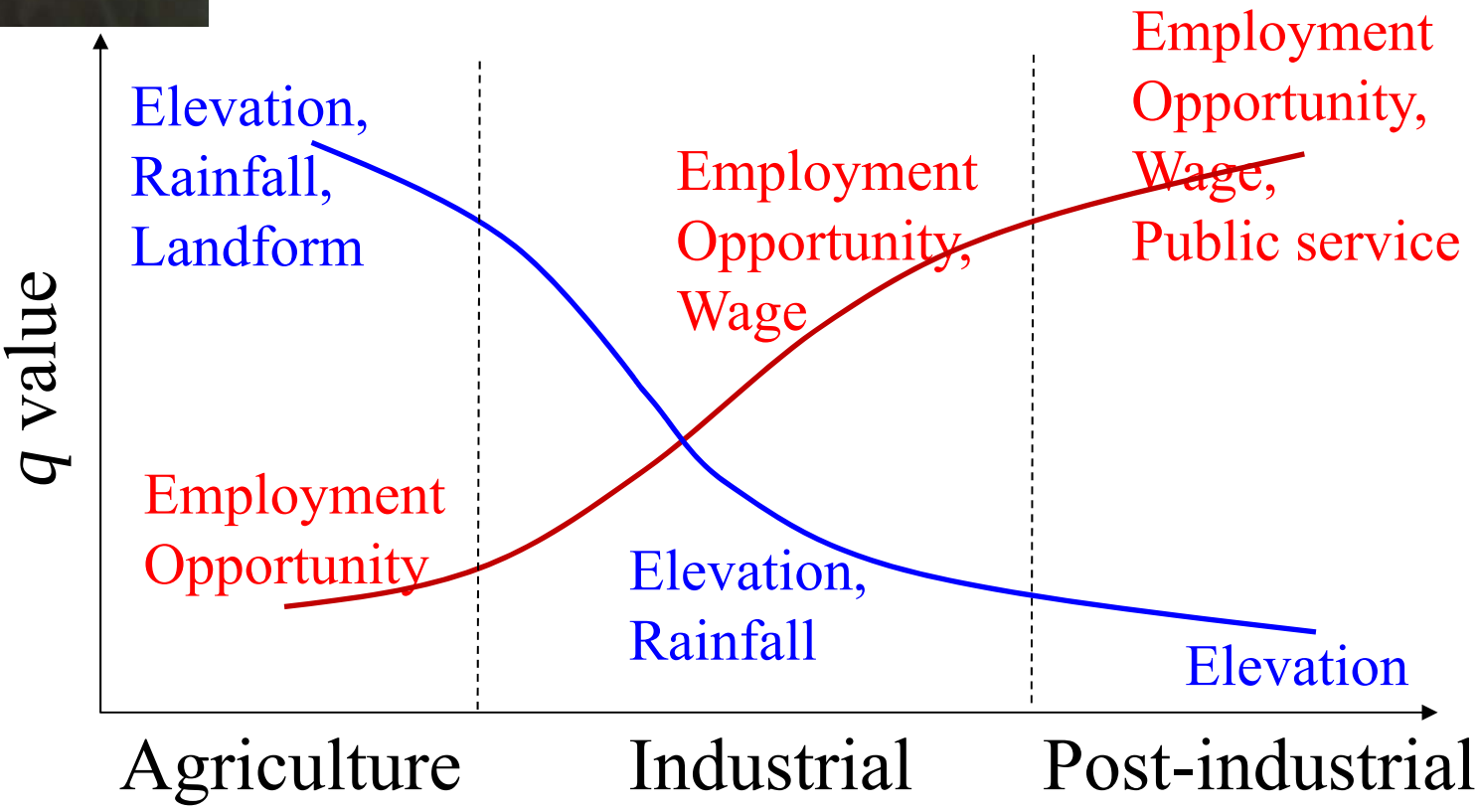
ESs=ecosystem services; Biop_Factor=biophysical factor; Clim_Factor=climate factor; Dev_Factor=development factor; EndStress_Fac=environmental stress factor; Ldeg_Fac= land degradation factor; Soceco_Fac = socioeconomic factor.



- Interaction btw biophysical and climatic factors on ES is greatest
- Biophysical factors very weakly interact development factors
- Climate factors very strongly interact socio-economic factors

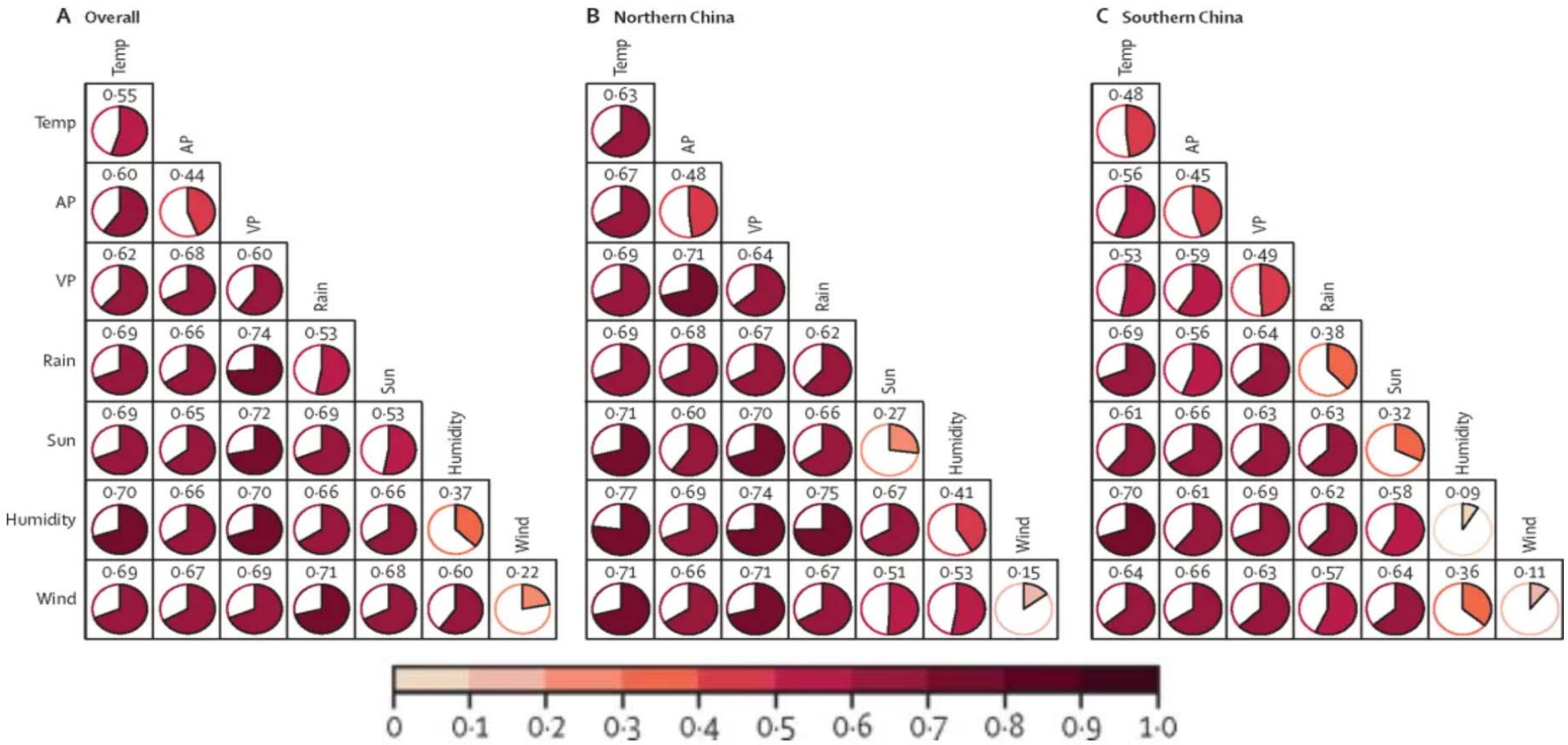
Sannigrahi et al. 2020. Responses of ecosystem services to natural and anthropogenic forcings: A spatial regression based assessment in the world's largest mangrove ecosystem. Science of the Total Environment 715: 137004.

Story 6. Determinants' Evolution of the Hu Line



Li JM et al. 2019. A Balanced development: Nature environment and economic and social power in China. Journal of Cleaner Production 210: 181-189.

Story 7. Meteorological Factors' Interaction on Virus

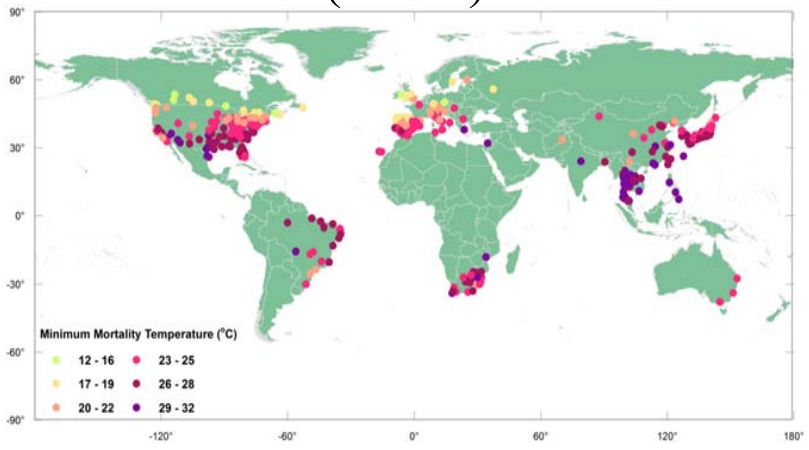


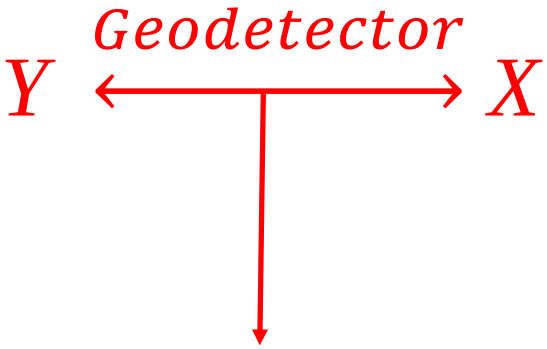
Interactive effects of paired meteorological factors on total virus-positive rates, by region
diagonal show *q*-statistics for the individual factors without interactions. Temp=temperature. AP=atmospheric pressure. VP=vapour pressure, Rain=rainfall. Sun=hours of sunlight. Humidity=relative humidity. Wind=wind speed.

Xu B et al. 2021. Seasonal association between viral etiologies of hospitalized acute lower respiratory infections and meteorological factors in China. Lancet Planetary Health 5: e154-63

Story 8. Spatial Goodness of Fit of MMT Prediction

Minimum Mortality Temperature (MMT)





AMT

78th % Temp

MFT

Table 3 Statistical indices between the three temperature indicators and MMT

	Annual mean temperature	78th percentile temperature	MFT
Pearson correlation	0.71	0.75	0.93
q statistics	0.56	0.57	0.83

The Geodetector *q* statistic reports that the MFT explains the 83% spatially stratified heterogeneity of the MMT, much higher than AMT(56%) and 78% temperature (57%).

Yin Q et al. 2019. Mapping the increased mortality temperature in the context of global change. Nature Communications 10: 4640.

Story 9. Geodetector +

Geodetector + Kriging

Liu YL et al. 2021. Geographical detector-based stratified regression kriging strategy for mapping soil organic carbon with high spatial heterogeneity. **Catena** 196 (2021): 104953.

Geodetector + GWR

Wu XX, et al. 2021. Natural and human environment interactively drive spread pattern of COVID-19: A city-level modeling study in China. **Science of the Total Environment** 756: 143343.

Geodetector + Rough set

Bai HX, Li DY, Ge Y, Wang JF, Cao F. 2021. Spatial rough set-based geographical detectors for nominal target variables. **Information Sciences**. https://doi.org/10.1016/j.ins.2021.12.019.

Geodetector + InSAR

Chen J, et al. 2022. Magnitudes and patterns of large-scale permafrost ground deformation revealed by Sentinel-1 InSAR on the central Qinghai-Tibet Plateau. **Remote Sensing of Environment** 268: 112778.

Geodetector + Google Earth Engine

Liu CL, et al. 2020. Land use/land cover changes and their driving factors in the Northeastern Tibetan Plateau based on Geographical Detectors and Google Earth Engine: A case study in Gannan Prefecture. **Remote Sensing** 2020, 12, 3139.

Geodetector + SWAT

Wang ZY, Cao JS. 2021. Spatial-temporal pattern study on water conservation function using the SWAT model. **Water Supply**. doi: 10.2166/ws.2021.127.

Geodetector + Random Forest

Zhou XZ, et al. 2021. Landslide susceptibility mapping using hybrid random forest with GeoDetector and RFE for factor optimization. **Geoscience Frontiers**. https://doi.org/10.1016/j.gsf.2021.101211.

Geodetector + Machine Learning

Wang Q, et al. 2021. Development of a new framework to estimate the environmental risk of heavy metal(loid)s focusing on the spatial heterogeneity of the industrial layout. **Environment International** 147: 106315

Geodetector + BHM

Wang ZS, et al. 2022. Measuring spatial nonstationary effects of POI-based mixed use on urban vibrancy using Bayesian spatially varying coefficients model. **International Journal of Geographical Information Science.** https://doi.org/10.1080/13658816.2022.2117363

Geodetector + ?

Statistics & Spatial Data

Spatial Stratified Heterogeneity (SSH)

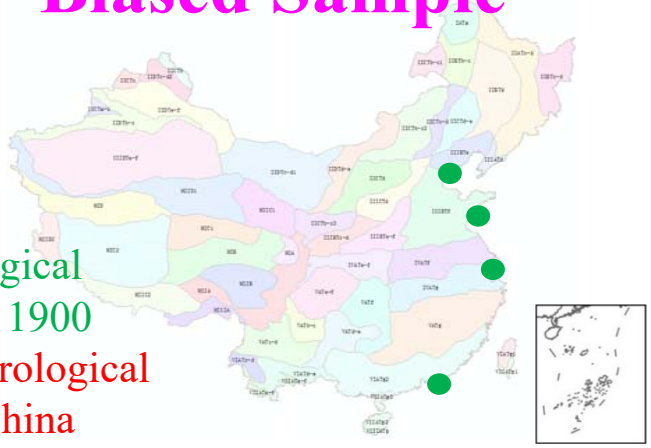
Causation of/by SSH

9 Stories

Discussion

Conclusion

Biased Sample



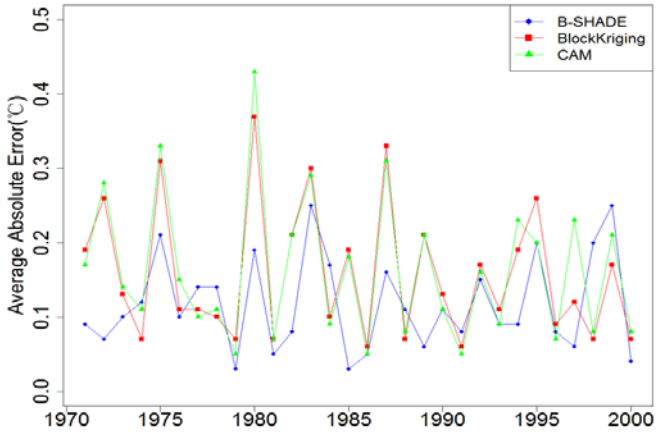
Meteorological stations in 1900 and meteorological zones in China

Confounded

	Treatment A			Treatment B		
	Works	Total	W/T	Works	Total	W/T
Small Stone	81	87	93%	234	270	87%
Large Stone	192	263	73%	55	80	69%
Overall	**273**	**350**	**78%**	**289**	**350**	**83%**

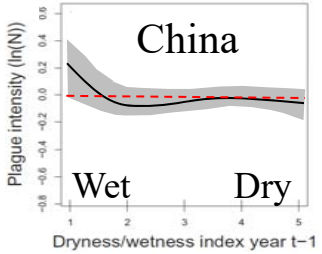
Charig CR, et al. 1986. *Br Med J (Clin Res Ed)* 292 (6524): 879–882.

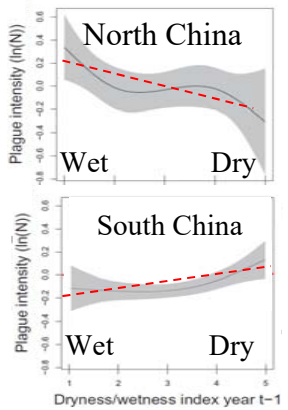
Biased Statistic



Wang JF, Xu CD, Hu MG, *et al*. 2014. *JGR* 119(1): 1-9.

Misleading CI





Xu L, *et al*. 2011. *PNAS*

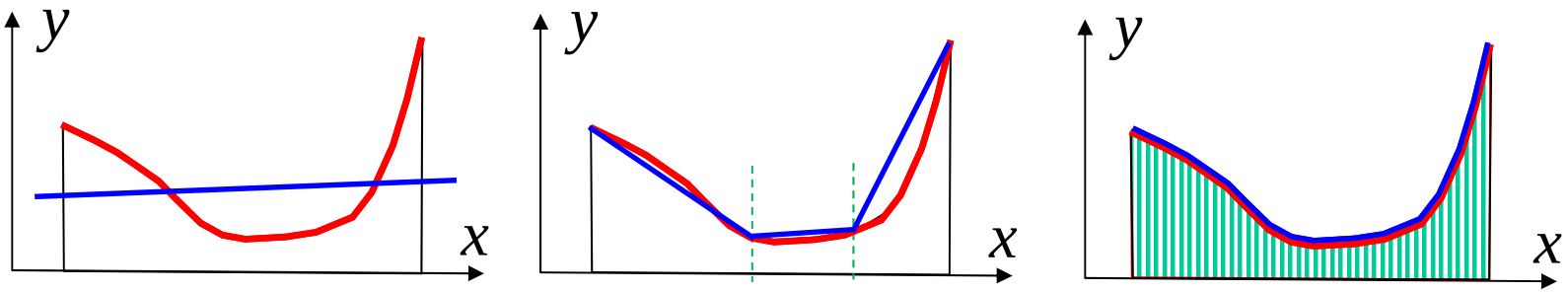
are sourced from: Spatial Stratified Heterogeneity

solution: find SSH so bias or confound, then Q/A

Wang JF, Zhang TL, Fu BJ. 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 7: 250-256.

Nonlinear Causation

Universe & Modelling



Global/BHM

Confounding

Simpson Paradox

Geodetector

中

Local/GWR

Overfitting

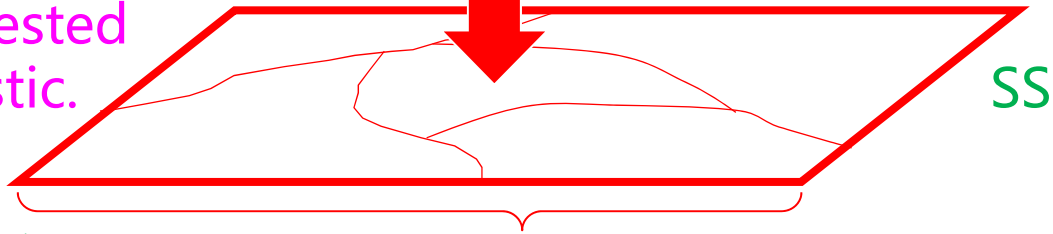
Biased Sample
Confounded Model
Biased Statistic
Misleading CI

Spatial Stratified Heterogeneity
is prevalent in the nature

Random
Spatial Pattern
Nonlinear Cause
General Interact

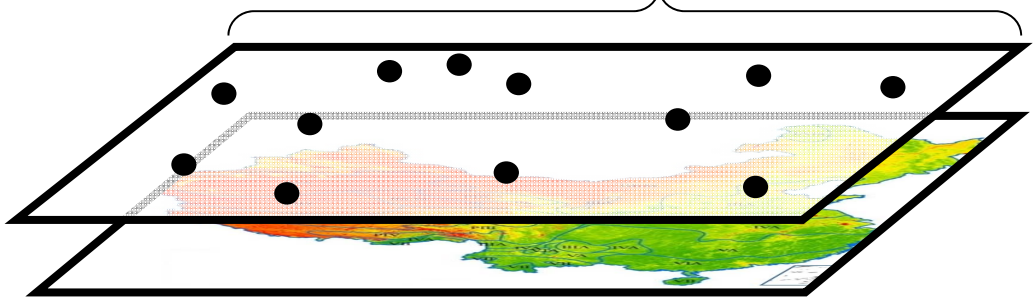
SSH should be tested before any statistic.

Solution:
modelling in strata



SSH is gold mine

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2}$$



www.geodetector.cn