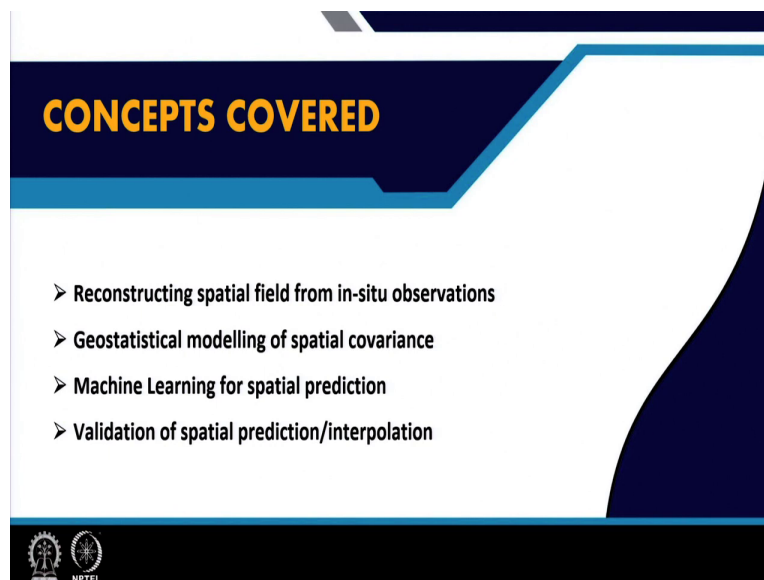


Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 03
Machine Learning for Discovering New Insights
Lecture - 23
Geostatistical modelling for mapping based on in-situ measurements

Hello, everyone. Welcome to lecture 23 of this course Machine Learning for Earth System Science. We are currently in module 3, where we are focusing on Machine Learning methods to like for discovering new insights related to various aspects of earth science. In today's lecture, we will be focusing on Geostatistical modelling for mapping based on in-situ measurements.

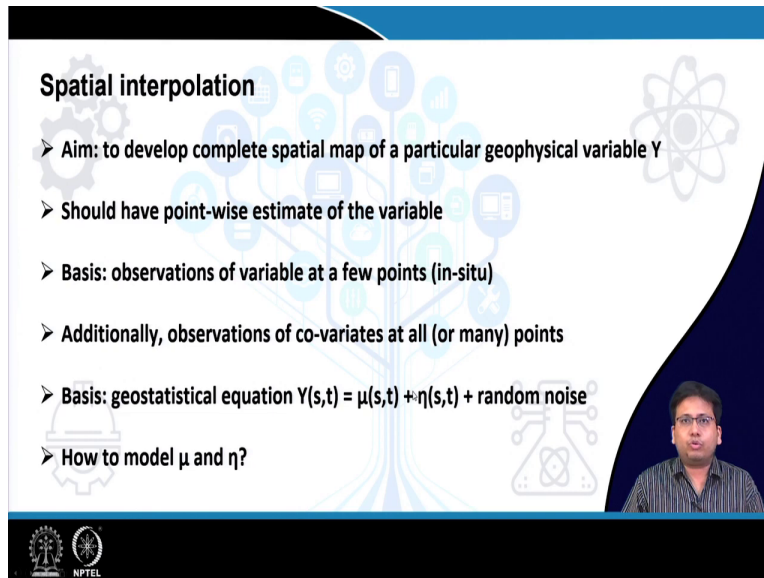
(Refer Slide Time: 00:52)



So, the concepts that we are going to cover today are first of all how to reconstruct a spatial field of any variable from in situ observations. Secondly, geostatistical modelling of spatial covariance like we have like this is a concept which we had studied in one of the very early lectures. So, today we will see some recent papers which have made use of that albeit in a more advanced way.

Thirdly, we will see how machine learning methods can be used for spatial predictions and how such spatial prediction or spatial interpolation can be validated against observations.

(Refer Slide Time: 01:30)



Spatial interpolation

- Aim: to develop complete spatial map of a particular geophysical variable Y
- Should have point-wise estimate of the variable
- Basis: observations of variable at a few points (in-situ)
- Additionally, observations of co-variables at all (or many) points
- Basis: geostatistical equation $Y(s,t) = \mu(s,t) + \eta(s,t) + \text{random noise}$
- How to model μ and η ?

The slide features a background with various icons related to geophysics and data science. A small video inset in the bottom right corner shows a man speaking. The NPTEL logo is visible in the bottom left corner.

So, first of all let us talk about the basic idea of spatial interpolation. The aim here is to develop the complete spatial map of a particular geophysical variable Y . By that I mean that I should have point wise estimate of the variable. You give me any point any location by specifying its exact coordinates like I should be able to give you the measurements of that of the variable which is of interest. It is not gridded observations it is like point wise observations.

But, on the what basis should I construct such a spatial field? The basis is like I may have some observations obtained from some sensors like say thermometers or barometers or whatever which are placed like at only a specific points. Like maybe like on top of some tower there is a like there is a like wind measurement device, maybe there is a particular weather station somewhere inside a city which measures the temperature at that particular place and so on.

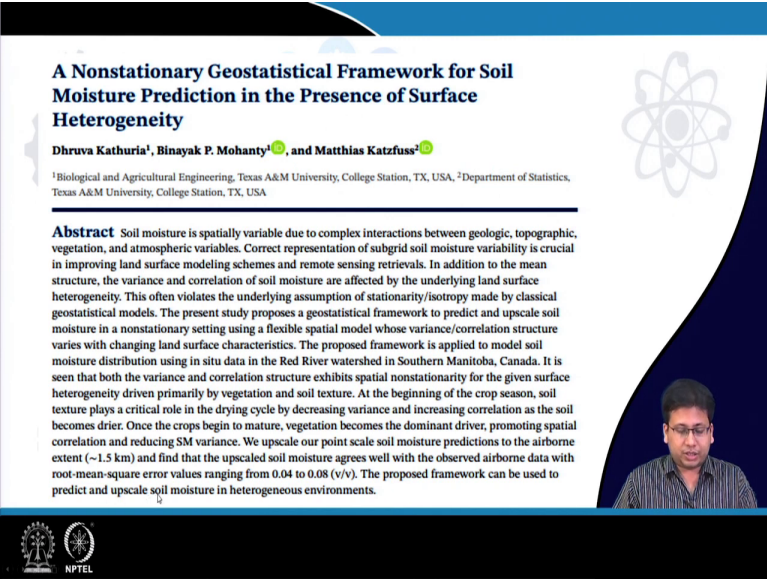
So, these are all what is known as in situ measurements that is I have point wise measurements at only a small number of points and additionally, I may also have observations of some covariates which are known to influence the geophysical variable Y which is of my interest and these

covariates may be the observation we may assume that these covariates are observed at all points or at least in many points.

Now, the basis for making such a spatial interpolation is the geostatistical equation which we had come across earlier. So, here like at that time we may have denoted the variable as X here we are denoting it by Y . So, it is a spatiotemporal variable. So, we divide it into the local component, the global component plus the noise. So, now, the question is how to model μ and η .

So, the different the research question basically based on to that how to get μ and η . So, some highlights of that or some basic outline to that we had discussed at that time when we considered a Bayesian hierarchical model using Gaussian process and so on. But, let us see like what are the challenges faced in real world problems where this is used and how researchers get around them.

(Refer Slide Time: 04:02)



A Nonstationary Geostatistical Framework for Soil Moisture Prediction in the Presence of Surface Heterogeneity

Dhruva Kathuria¹, Binayak P. Mohanty¹, and Matthias Katzfuss²

¹Biological and Agricultural Engineering, Texas A&M University, College Station, TX, USA, ²Department of Statistics, Texas A&M University, College Station, TX, USA

Abstract Soil moisture is spatially variable due to complex interactions between geologic, topographic, vegetation, and atmospheric variables. Correct representation of subgrid soil moisture variability is crucial in improving land surface modeling schemes and remote sensing retrievals. In addition to the mean structure, the variance and correlation of soil moisture are affected by the underlying land surface heterogeneity. This often violates the underlying assumption of stationarity/isotropy made by classical geostatistical models. The present study proposes a geostatistical framework to predict and upscale soil moisture in a nonstationary setting using a flexible spatial model whose variance/correlation structure varies with changing land surface characteristics. The proposed framework is applied to model soil moisture distribution using in situ data in the Red River watershed in Southern Manitoba, Canada. It is seen that both the variance and correlation structure exhibits spatial nonstationarity for the given surface heterogeneity driven primarily by vegetation and soil texture. At the beginning of the crop season, soil texture plays a critical role in the drying cycle by decreasing variance and increasing correlation as the soil becomes drier. Once the crops begin to mature, vegetation becomes the dominant driver, promoting spatial correlation and reducing SM variance. We upscale our point scale soil moisture predictions to the airborne extent (~1.5 km) and find that the upscaled soil moisture agrees well with the observed airborne data with root-mean-square error values ranging from 0.04 to 0.08 (v/v). The proposed framework can be used to predict and upscale soil moisture in heterogeneous environments.

So, let us focus on this paper which came out in 2019, A stationary Geostatistical Framework for Soil Moisture Prediction in the Presence of Surface Heterogeneity. This concept of this heterogeneity this is the big point here. So, soil moisture is spatially is spatially variable due to complex interactions between geologic topographic vegetation and atmospheric variables.

Correct representation of subgrid soil moisture variability is crucial in improving land surface modeling schemes and remote sensing retrievals. Note the word subgrid here, I am not talking

about gridded observations. Earlier I may have talked about like when we are discussing climate networks and so on, we repeatedly mentioned gridded observations where the entire surface is divided into grids and we like represent each grid by one value.

So, it is like we have a discrete number of measurements to be made, but in this case it is like it might be continuous, that is, instead of restricting ourselves to any grid we are going to point wise measurements in this case. In addition to the mean structure the variance and correlation of soil moisture are affected by the underlying land surface heterogeneity.

This often violates the underlying assumption of stationarity or isotropy made by classical geostatistical models. This is a the very important issue like we had the stationarity is one concept which we had discussed earlier. And, when we do like linear the usual spatial interpolation such as kriging like one of the main concept or one of the main assumptions there is that of spatial stationarity.

But, in this case what if the stationarity does not hold? The present study proposes a geostatistical framework to predict and upscale soil moisture in a nonstationary setting using a flexible spatial model whose variance or correlation structure varies with changing land surface characteristics; that is, in different parts of the region the land surface characteristics are different.

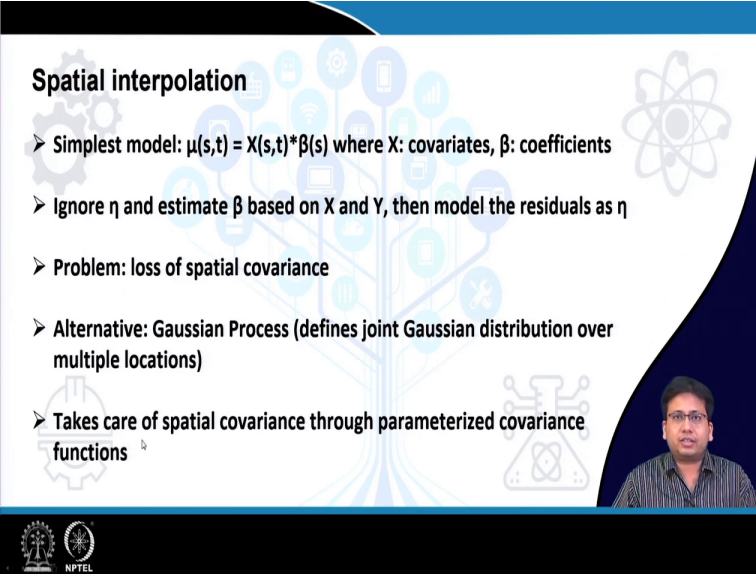
The proposed framework is applied to model soil moisture distribution using in situ data in the red river watershed of South Manitoba in Canada. It is seen that both the variance and correlation structure exhibits spatial nonstationarity for the given surface heterogeneity primarily due to vegetation and soil texture. At the beginning of crop season, soil texture plays a critical role in drying in the drying cycle by decreasing variance and increasing correlation as the soil becomes drier.

Once the crops begin to mature, vegetation becomes the dominant driver, promoting spatial correlation and reducing soil moisture variance. So, like the point that we are making here is that the variance structure the like if you consider the soil moisture in location s_1 and that in location s_2 the variance the covariance between them is actually a function of time. In different days of

the year depending on the like depending on the cropping and so on like this variance relation is found to be changing.

We upscale our point scale soil moisture predictions to the airborne extent and find that the upscale soil moisture agrees well with the observed airborne data with are root mean square error values ranging from 0.04 to 0.08. The proposed framework can be used to predict and upscale soil moisture in heterogeneous environments.

(Refer Slide Time: 08:16)



Spatial interpolation

- Simplest model: $\mu(s,t) = X(s,t) * \beta(s)$ where X: covariates, β : coefficients
- Ignore η and estimate β based on X and Y, then model the residuals as η
- Problem: loss of spatial covariance
- Alternative: Gaussian Process (defines joint Gaussian distribution over multiple locations)
- Takes care of spatial covariance through parameterized covariance functions

The slide features a background with various icons related to data science and statistics. A small video feed of a presenter is visible in the bottom right corner of the slide area. The NPTEL logo is at the bottom left.

So, the last part is related to the validation. So, we are able to make point wise predictions, but how to validate them? So, we obviously, we cannot have like we cannot have measurements at every point. So, what is done is like they have airborne predictions like may be through remote sensing or something like that at like very high spatial resolution and the measurement the estimates that are made using our model are going to be validated against them.

So, now when we come to spatial interpolation, so, earlier I was mentioned that the main challenge here is to mathematically model the local component μ and the global component η . So, the simplest way to go about it the way people used to do it in classically the that is the people in let us say in the fields of petroleum engineering and so on where the geostatistics is very important.

So, what they did is they this the local component $\mu(s, t) = X(s, t) * \beta(s)$ where X like these are basically the covariates. So, this you can say is a vector of covariates at that location and time point and $\beta(s)$ this is a vector of the corresponding coefficients. So, they basically model the local component in terms of different covariates.

And, now they first ignore the η part altogether and just try to estimate this β based on X and Y , that is they framed it as a they temporarily forgot the η and just like focused on Y , X and β as a linear interpolation problem I mean sorry linear regression problem and solved β . So, obviously, the matching will not be perfect. So, there will be some residuals.

So, like this way you will get some value which will be differ which will not even if you are trying to fit this value to Y it will not fit directly some error will be there. So, that residual that is to considered as η . In reality it is actually η plus the random noise, but let us forget the random noise and just say that the residual is η . And, then we the like then we try to fix some kind of a model be it Gaussian process or something like that on the η which are obtained from the different locations at any point of time or at different points of time.

So, it is like you first forget η , estimate μ , then whatever is the residual use interpret that as η and then fit some model on it. The problem with this approach is a loss of spatial covariance. For obvious reasons when you are these β is specific two locations that is like you are basically the μ component is like it is being measured independently at each location while that is generally not the case.

Even the although μ is the local component it is considered to be a property of a location, but even that is expected to be spatially smooth it cannot it is not expected to vary significantly from one location to another. So, you cannot really or it is a big like we are making a big approximation when we are treating the μ is at every location independently.

However, that is how classical geostatistics used to work. The result was the loss of spatial covariance; the covariance structure of the field especially in of a heterogeneous field was often lost out on. And, alternative is the least of course, the Gaussian process which we have discovered or which we have discussed in the one of the earlier lectures. So, this takes care of spatial covariance through parameterized covariance function.

(Refer Slide Time: 12:09)

Modelling with Gaussian Processes

$$\mathbf{y} = (y(s_1), \dots, y(s_n)) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) = (2\pi)^{-n/2} |\mathbf{C}|^{-1/2} \exp(-(1/2)(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{y} - \boldsymbol{\mu})),$$

$$\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2} (\log |\mathbf{C}(\boldsymbol{\theta})| + (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))^T \mathbf{C}^{-1}(\boldsymbol{\theta}) (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})) + n \log(2\pi))$$

MAXIMUM-LIKELIHOOD ESTIMATE:
 $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{C}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{y}.$

M surface control processes

$$\mathbf{e}(s) = \sum_{j=1}^M w_j(s) \mathbf{e}_j(s) \quad w_j(s) = \frac{\exp(\mathbf{X}(s)^T \boldsymbol{\alpha}_j)}{\sum_{i=1}^M \exp(\mathbf{X}(s)^T \boldsymbol{\alpha}_i)}$$

$$\text{Cov}(\mathbf{e}(s_1), \mathbf{e}(s_2)) = \text{Cov} \left(\sum_{j=1}^M w_j(\mathbf{X}(s_1)) \mathbf{e}_j(s_1), \sum_{j=1}^M w_j(\mathbf{X}(s_2)) \mathbf{e}_j(s_2) \right)$$

$$= \sum_{j=1}^M w_j(\mathbf{X}(s_1)) w_j(\mathbf{X}(s_2)) \mathbf{C}_j(s_1, s_2)$$

$$= \mathbf{C}(s_1, s_2, \mathbf{X}(s_1), \mathbf{X}(s_2))$$

$$\mathbf{C}_k^{\text{low}}(s) = \sum_{j=1}^M \frac{\exp(\alpha_{j1} + (x_k^{(0,1)}) \alpha_{jk})}{\sum_{i=1}^M \exp(\alpha_{i1} + (x_k^{(0,1)}) \alpha_{ik})} \mathbf{C}_j(s)$$

$$\mathbf{C}_k^{\text{dominant}}(s) = \sum_{j=1}^M \frac{\exp(\alpha_{j1})}{\sum_{i=1}^M \exp(\alpha_{i1})} \mathbf{C}_j(s)$$

$$\mathbf{C}_k^{\text{high}}(s) = \sum_{j=1}^M \frac{\exp(\alpha_{j1} + (x_k^{(0,9)}) \alpha_{jk})}{\sum_{i=1}^M \exp(\alpha_{i1} + (x_k^{(0,9)}) \alpha_{ik})} \mathbf{C}_j(s)$$

Kathuria et al, 2019

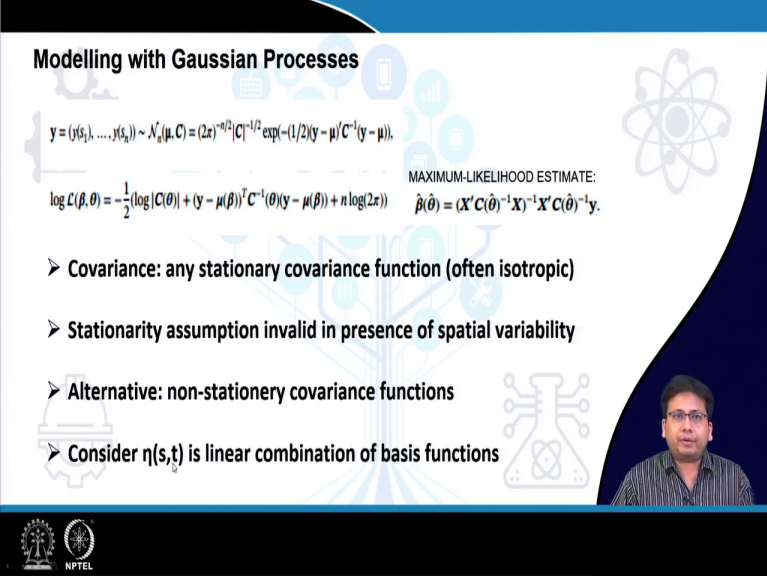
So, like. So, so first let us consider that there are n locations s_1, s_2, \dots, s_n and we are considering the y at all of them. So, now, the instead of considering them individually we consider they as a joint distribution and like if you as you remember in case of Gaussian process like if you consider any subset of variables their joint distribution is always Gaussian.

So, in this case also like when we are considering these n the observations of y at these n points, their joint distribution also follows this kind of a Gaussian distribution with mean vector which we denote by this μ . So, this μ is actually an it is an n -dimensional vector and the covariance which is C which is an $n \times n$ matrix.

So, now, we can like estimate the μ and C through maximum likelihood like provided we have the provided we are considering these n observation as from the those locations where we do have the in situ observations. So, from that we can estimate the μ and the C . Now, this μ if you like you may want to express that $\mu = X\beta$ the way we are saying earlier and so, basically the these like we can get an maximum likelihood estimate of both β and C .

Or if you want we like we can put some kind of a like a covariance structure on C or a variance function on C and do something with it.

(Refer Slide Time: 13:45)



Modelling with Gaussian Processes

$y = (y(s_1), \dots, y(s_n)) \sim \mathcal{N}_n(\mu, C) = (2\pi)^{-n/2} |C|^{-1/2} \exp(-(1/2)(y - \mu)^T C^{-1} (y - \mu))$

$\log \mathcal{L}(\beta, \theta) = -\frac{1}{2} (\log |C(\theta)| + (y - \mu(\beta))^T C^{-1}(\theta) (y - \mu(\beta)) + n \log(2\pi))$

MAXIMUM-LIKELIHOOD ESTIMATE:
 $\hat{\beta}(\hat{\theta}) = (X^T C(\hat{\theta})^{-1} X)^{-1} X^T C(\hat{\theta})^{-1} y$

- Covariance: any stationary covariance function (often isotropic)
- Stationarity assumption invalid in presence of spatial variability
- Alternative: non-stationary covariance functions
- Consider $\eta(s, t)$ is linear combination of basis functions

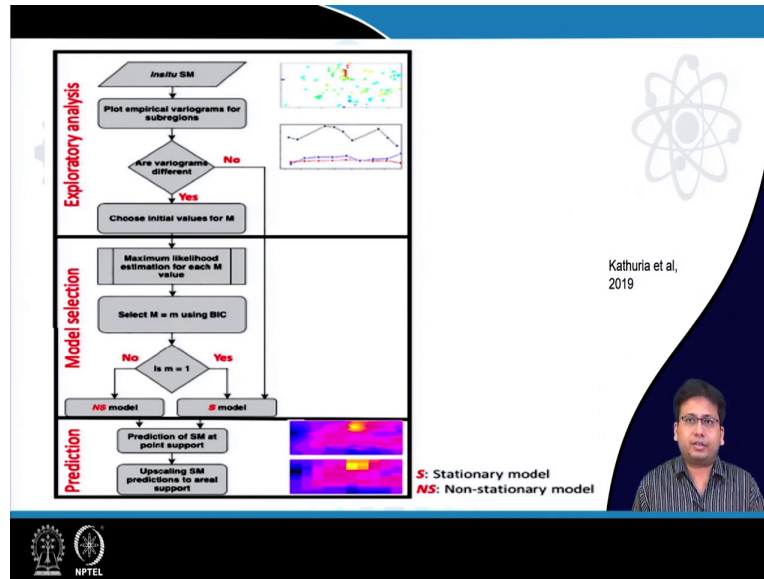
The like we can one that is what we many people often do is they simply consider any stationary covariance function like when we are discussing about Gaussian processes we have already talked about covariance function and like one like simple class of covariance functions is the like the isotropic ones where it is like consider that the basically the covariance is like I mean the variance is equal for in all cases like for all pairs of locations.

Now, the stationary or it can it need not be isotropic of course, it can be any stationary. By that I mean the covariance of the observations at any two locations is only a function of the distance between their between those two locations that is the definition of stationarity. But, if spatial variability is there especially in the case of the data which the authors are considering in this paper then this stationarity assumption is no longer valid.

So, the alternative is to develop some non-stationarity non-stationary covariance functions that is instead of like C being a stationary we have to somehow make it non-stationary. And, the one way to do it is we can this $\eta(s, t)$ this the η component which we did not consider so far like in the statistical like neither in the classical geostatistical sense nor in the with the Gaussian process the way it is handled so far we have considered the η so far.

Now, we can consider this η like as a linear combination of some certain basis functions. This is the way we where we will bring the heterogeneity in.

(Refer Slide Time: 15:36)



So, this is the rough approach which has been followed in this paper. So, the approach is as follows. First you have the in-situ measurements at some of the soil moisture at some locations at some isolated locations. So, based on that you plot the empirical variograms for the sub regions that is you plot the how the covariance varies as a the correlation between any two locations varies as a function of the distance and, you do it for different sub-regions.

The you the whole region you divide into several sub-regions in each region you plot the variogram like that and then see are the variograms different in the different region or not. If they are not the like if there is no difference; that means, it is a; it is the like the same region it is a homogeneous case. In that case you can just go to the like the statistic the classical approaches.

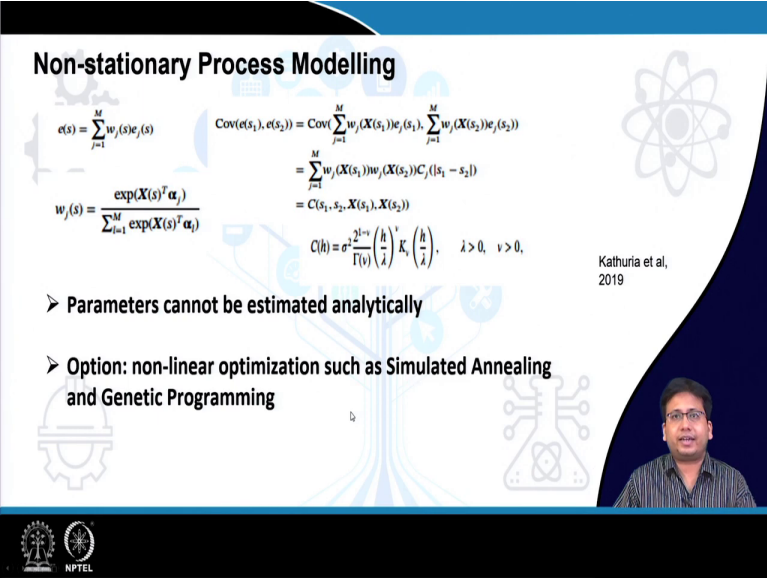
But, if the variograms are found to be different then it is necessary to like basically do something like dividing the whole region into sub-regions maybe not explicitly, but, implicitly by considering the drivers; drivers of what? Drivers of what causes the this kind of spatial heterogeneity. So, like we may not know how many drivers should be there initially. So, let us say let us say let us denote the number of such drivers as M .

So, for some like you have to play around a bit with the value of the M to know the know its optimal value. It is like somewhat like in case of clustering algorithms the way we play around with the value of k . So, for every value of M we have to estimate the model parameters using some process which we will see in the later slides.

And, once you have like once you have fit at a particular model then you have to validate it, you have to do the model evaluation if like if that works then good if not then you have to like re-evaluate the model and this process has to go on till we are able to do a good job of the model validation.

And, once that had once you have been able to like develop the model then what you can do is we can you develop the like you can predict the soil moisture like at point support that is for any given point which is of interest you can make the like point wise prediction of the soil moisture or the target variable in general. And, then you can like for validation purposes if it is necessary to upscale it back to the grid level, then you can do that also, ok.

(Refer Slide Time: 18:23)



Non-stationary Process Modelling

$$e(s) = \sum_{j=1}^M w_j(s) e_j(s)$$

$$w_j(s) = \frac{\exp(X(s)^T \alpha_j)}{\sum_{i=1}^M \exp(X(s)^T \alpha_i)}$$

$$\text{Cov}(e(s_1), e(s_2)) = \text{Cov}\left(\sum_{j=1}^M w_j(X(s_1)) e_j(s_1), \sum_{j=1}^M w_j(X(s_2)) e_j(s_2)\right)$$

$$= \sum_{j=1}^M w_j(X(s_1)) w_j(X(s_2)) C_j(|s_1 - s_2|)$$

$$= C(s_1, s_2, X(s_1), X(s_2))$$

$$C(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{h}{\lambda}\right)^{\nu} K_{\nu}\left(\frac{h}{\lambda}\right), \quad \lambda > 0, \quad \nu > 0,$$

➤ Parameters cannot be estimated analytically

➤ Option: non-linear optimization such as Simulated Annealing and Genetic Programming

Kathuria et al, 2019

So, now comes to the question the how to do this kind of a non-stationary process modelling. So, this $e(s)$ this is the here I am borrowing the notation from the paper, this e is what we have been calling as η the global component so to say. So, let us say that at any given location s the $e(s)$ is

like a linear combination of different drivers like this. So, just like the μ we considered as a linear combination of the different covariates, here we are considering that there are various drivers which are like denoted by this $e_1, e_2, e_3, \dots, e_j$ etcetera and each of them have their weights.

Now, the weights of the different drivers are calculated in this like according to this kind of a relation by the paper. So, this X and α these are like X are again those covariates which were used for μ also and α is another set of parameters. So, these just like the parameters of called β the parameter for μ had to be estimated. So, these α parameters also will have to be estimated for this e ok.

Now, it can be shown that the covariance between the this η at any two locations like you can express it in this particular way. So, like some kind of function has to be developed or has to be chosen. So, there are again some standard covariance functions to be used.

Now, these the problem with the estimation of these parameters β, α etcetera is that we will not be able to estimate them analytically; like in case of linear interpolation we could we had the β parameters which we could for which we had a closed form expression in this case unfortunately we will not have it. So, the option is to go for non-linear optimization or numerical techniques such as simulated annealing and genetic programming.

(Refer Slide Time: 20:22)

Non-stationary Process Modelling

➤ Prediction at new points can be carried out using conditional Gaussian distribution

$$\begin{pmatrix} y(S^p) \\ y \end{pmatrix} \sim \mathcal{N}_{n+1} \left(\begin{pmatrix} \hat{\mu}(S^p) \\ \hat{\mu}(S) \end{pmatrix}, \begin{pmatrix} \hat{C}(S^p, S^p) & \hat{C}(S^p, S) \\ \hat{C}(S, S^p) & \hat{C}(S, S) \end{pmatrix} \right)$$

$$y(S^p) | y \sim \mathcal{N}_{n_p}(\mu_{S^p|y}, C_{S^p|y}),$$

$$\mu_{S^p|y} = \hat{\mu}(S^p) + \hat{C}(S^p, S)(\hat{C}(S, S))^{-1}(y - \hat{\mu}(S)),$$

$$C_{S^p|y} = \hat{C}(S^p, S^p) - \hat{C}(S^p, S)(\hat{C}(S, S))^{-1}\hat{C}(S, S^p),$$

Kathuria et al, 2019

NPTEL

Now, once you have estimated the parameters now you are in a position to do the prediction at new points using conditional Gaussian distribution. So, let us say that like S^P is a point where you want to make the prediction of y which is the soil moisture and you they like you already have it is measurement at n other locations. So, those all those n measurements they are clubbed into this variable y .

So, now, remember Gaussian process whenever we have any collection of variables their joint distribution again follows a Gaussian distribution. So, these so, now, we have total $n + 1$ variables 1 here and n here. So, they are this subset of $n + 1$ variables is again going to have a joint Gaussian distribution which is given like whose parameters can be broken down like this so, ok.

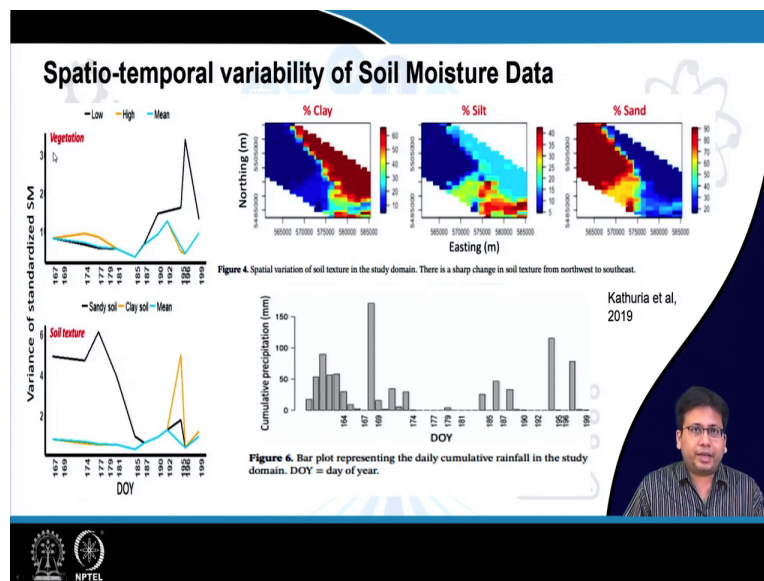
So, it is like the this is the $(n + 1)$ -dimensional μ vector which again can be divided into $1 + n$ and this is the $(n + 1) \times (n + 1)$ covariance matrix which also can be divided into blocks, like this is 1×1 , this is $1 \times n$, this is $n \times 1$ and this is $n \times n$. So, the like basically the once you are able to estimate these mean and covariance values then you like you for the variable which you are interested in the $y(S^P)$ you can say that it is conditional distribution is again a Gaussian distribution with particular μ and particular covariance.

So, that is. So, how you come down from this joint distribution of $y(S^P)$ and y to this conditional distribution of $y(S^P)|y$ this is left as an exercise to you. So, I think when we had come across Gaussian processes we had we at that time also I may have mentioned this exercise I hope you have done it by now.

So, the question is what about what are these the μ and C parameters of this conditional distribution. So, both of them can be estimated from the μ and C parameters which we had in the original Gaussian distribution. And, it should be rather easy for you to calculate the values of these the these blocks of C . So, you can do it easily using this covariance function which we had discussed here, ok.

And, the for the conditional distribution the mean and the covariance parameters like for them we have this kind of conditional values.

(Refer Slide Time: 23:16)



And now coming to the results the in this paper they have done extensive experiments. So, like they were they had mentioned two drivers the vegetation and the soil texture. These are the drivers of the e or the η which they are talking about this M . So, in this case $M = 2$ as they have two main drivers vegetation and soil texture and like and.

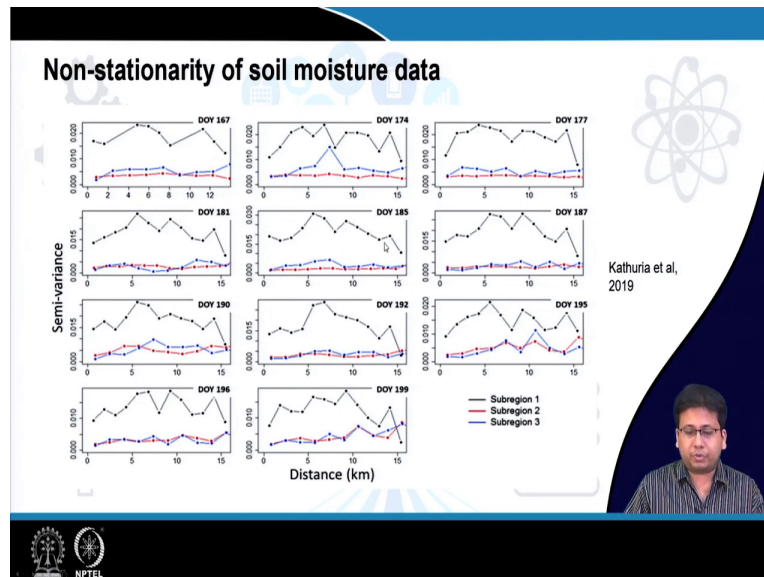
So, like basically they have shown how that these how these things like vary over the days of the year and so, DOY here stands for a Day Of the Year and in fact, for different kinds of soil they have actually plotted this to show how different the behaviour is and like this is the study area which they have considered.

So, like this is like you can say these are the longitudes and these are the latitudes. So, this is the north – south direction, this is the east – west direction. So, here like they are basically showing how the different properties of the soil is changing from one location to another. So, like basically the soil consists of these three components clay, silt and sand. So, here they are showing the percentage of the three components in different locations and so, that way you can see some clear regions.

So, here is in this part there is a region which is mostly sandy along this part like again a diagonalish component which is mostly clayey and then there is one region towards the like we this should be the south east which is more mixed and which has a significant amount of silt also, ok. And, and remember that apart from these drivers we had considered the covariates also for the μ part and the most important covariate they have considered here is the cumulative rainfall precipitation.

So, we from common knowledge we can understand that the rain the soil moisture is directly related to the rainfall. Now, rainfall of course, has a specific season and like this is like across the different days of the year this is how the rainfall is found to be varying.

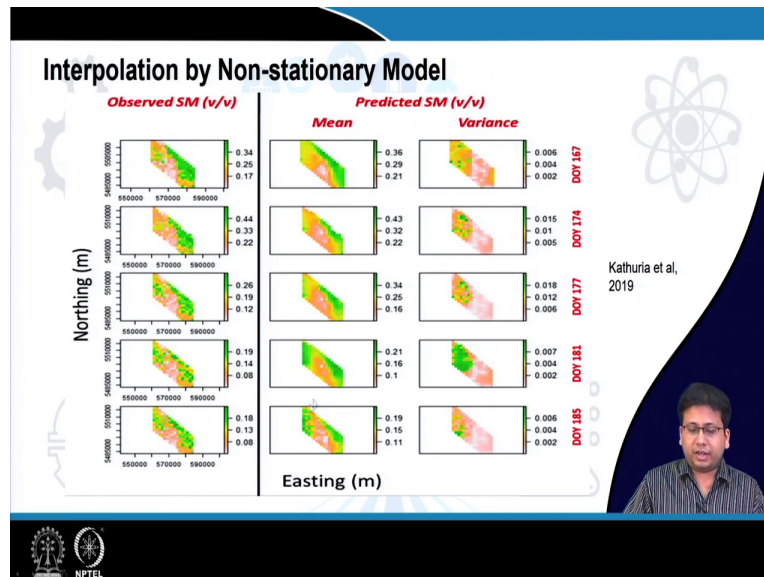
(Refer Slide Time: 25:32)



Now, also to show the non-stationarity of the soil moisture data as we talked about like it is necessary to construct the variograms in different regions. So, here they have divided the region as shown here into subregions 1, 2 and 3 and in each case like they have actually carried out this kind of like variogram analysis to show the variation.

And, so, this is to basically to understand that the data is non stationary non stationary by nature. So, these different plots are different corresponding to different days of the year.

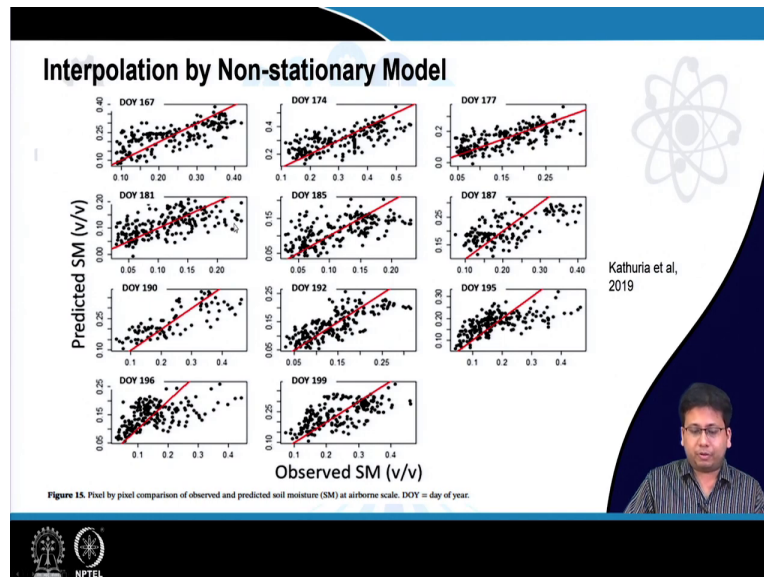
(Refer Slide Time: 26:11)



And, here well the so, here basically they have shown the results of the interpolation. So, these are the observed values of the soil moisture it is plotted over the region and these are the predicted values.

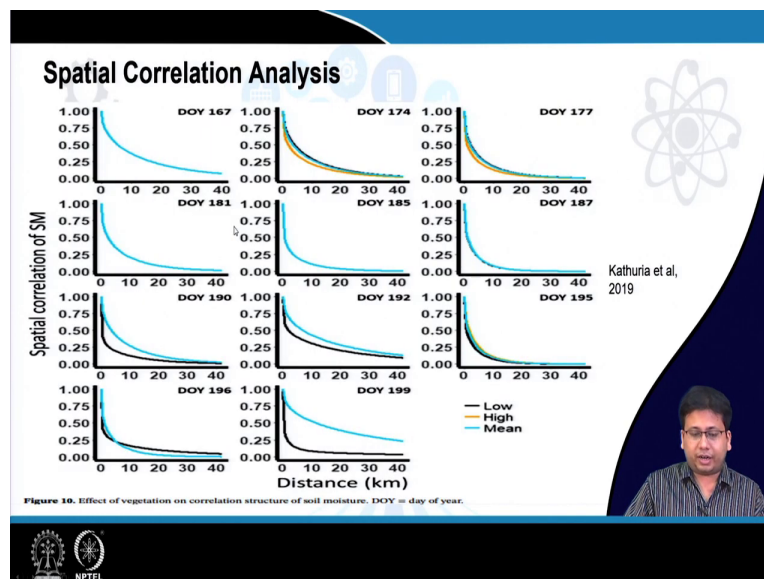
So, the mean the when they are predicting the soil moisture because it is coming with the predictions are being done by a conditional Gaussian as shown here. So, they need to the mean is the point wise estimate, but they also need to focus on the variance. So, they have plotted both the mean and the variance here and it is possible to will actually correlate this the point wise estimate the mean map with the actual observations and calculate some errors based on that.

(Refer Slide Time: 26:52)



So, these are the error plots of the predicted values versus the observed values and you in most cases you can see a more or less like a strong positive correlation. That is, you can in all cases you can see it is like the this line the one correlation line almost all the points are like around located around it.

(Refer Slide Time: 27:15)



And, so, similarly we can have the variogram analysis of the estimated values. So, like in on like so, like in like here we on the original data we had carried out this kind of analysis. So, similar analysis on can be carried out where they are estimating the variance as a function of distance ok.

(Refer Slide Time: 27:41)

A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration

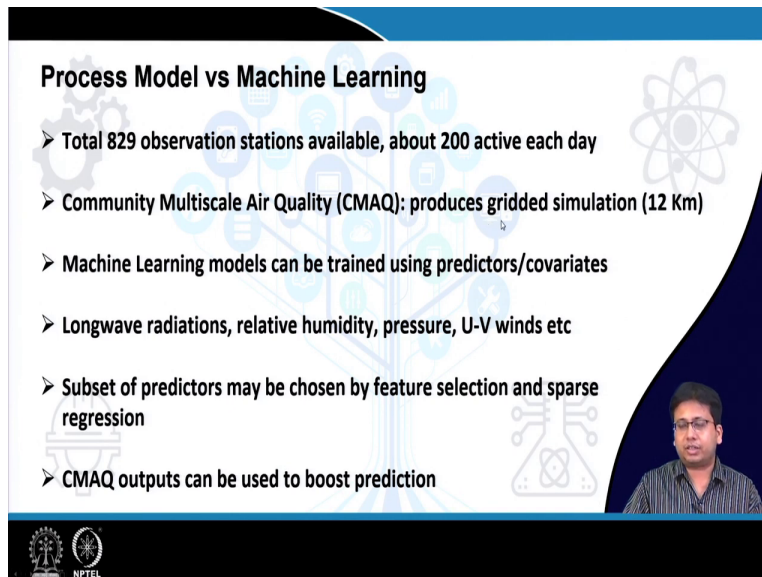
Veronica J. Berrocal^{a,*}, Yawen Guan^b, Amanda Muyskens^c, Haoyu Wang^d, Brian J. Reich^e, James A. Mulholland^f, Howard H. Chang^g

ABSTRACT

A typical challenge in air pollution epidemiology is to perform detailed exposure assessment for individuals for which health data are available. To address this problem, in the last few years, substantial research efforts have been placed in developing statistical methods or machine learning techniques to generate estimates of air pollution at fine spatial and temporal scales (daily, usually) with complete coverage. However, it is not clear how much the predicted exposures yielded by the various methods differ, and which method generates more reliable estimates. In this paper, we aim to address this gap by evaluating a variety of exposure modeling approaches, comparing their predictive performance. Using $PM_{2.5}$ in year 2011 over the continental U.S. as a case study, we generate national maps of ambient $PM_{2.5}$ concentration using: (i) ordinary least squares and inverse distance weighting; (ii) kriging; (iii) statistical downscaling models, that is, spatial statistical models that use the information contained in air quality model outputs; (iv) land use regression, that is, linear regression modeling approaches that leverage the information in Geographical Information System (GIS) covariates; and (v) machine learning methods, such as neural networks, random forests and support vector regression. We examine the various methods' predictive performance via cross-validation using Root Mean Squared Error, Mean Absolute Deviation, Pearson correlation, and Mean Spatial Pearson Correlation. Additionally, we evaluated whether factors such as season, urbanicity, and levels of $PM_{2.5}$ concentration (low, medium or high) affected the performance of the different methods. Overall, statistical methods that explicitly modeled the spatial correlation, e.g. universal kriging and the downscale model, outperform all the other exposure assessment approaches regardless of season, urbanicity and $PM_{2.5}$ concentration level. We posit that the better predictive performance of spatial statistical models over machine learning methods is due to the fact that they explicitly account for spatial dependence, thus borrowing information from neighboring observations. In light of our findings, we suggest that future exposure assessment methods for regional $PM_{2.5}$ incorporate information from neighboring sites when deriving predictions at unsampled locations or attempt to account for spatial dependence.

And, now we come to another paper where like basically the idea is to compare the statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration. So, like $PM_{2.5}$ as we know is like it is a like it is a kind of particle which is used to measure the air quality.

(Refer Slide Time: 28:08)



Process Model vs Machine Learning

- Total 829 observation stations available, about 200 active each day
- Community Multiscale Air Quality (CMAQ): produces gridded simulation (12 Km)
- Machine Learning models can be trained using predictors/covariates
- Longwave radiations, relative humidity, pressure, U-V winds etc
- Subset of predictors may be chosen by feature selection and sparse regression
- CMAQ outputs can be used to boost prediction

The slide features a background with faint icons of a gear, a network, and a molecular structure. A small inset video of a speaker is visible in the bottom right corner. The NPTEL logo is at the bottom left.

So, again the point the task is that like they have 829 observation stations for air quality out of which only 200 are active each day that is they have sparse in situ measurements. So, the what they want to do is they want to develop a pan US map of this air quality that is a point wise map of air quality based on only 829 observation stations which are we like we can say are in situ measurements. And, not only is the total number quite low 829 compared to such a vast country as US, but only on any given day only 200 of them are active.

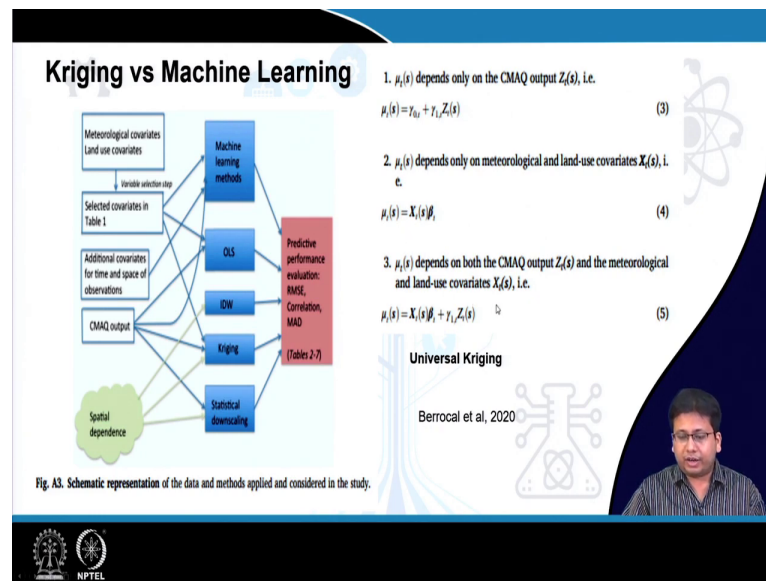
Now, Community Multiscale Air Quality or CMAQ this is something of a something like a process model. So, earlier also we have talked about process based simulation models. So, this for air quality this is a CMAQ is a process based simulation models which produces gridded simulations at high resolution that is 12 kilometre by 12 kilometres. So, this is the like a traditional way of doing the like building the map, but again this is based on purely based on physics.

So, an alternative is to use machine learning models say random forest then support vector regression, neural networks etcetera where like at every location we can depend on certain predictors or covariates such as longwave radiations, relative humidity, pressure, the horizon the east-west and north-south winds etcetera all of these are known to affect the air the air quality

and $PM_{2.5}$ in some way or the other. So, like we can use all these covariates linearly or non-linearly through machine learning models to make the predictions.

Now, out of all these predictors not all of them are of course, going to be useful. So, like we can choose the most useful ones through like feature selection such as sparse regression say things like a lasso regression and so on which like we have discussed these earlier also where we have a initially have a large pool of predictors from which we identify or the pull out the most important ones.

(Refer Slide Time: 30:18)



Now, the CMAQ outputs can actually be used to boost the predictions. So, like these like. So, basically you have the meteorological covariates and as well as the land use covariates. So, now, you can use something like a variable selection as we mentioned to like select only a specific covariates.

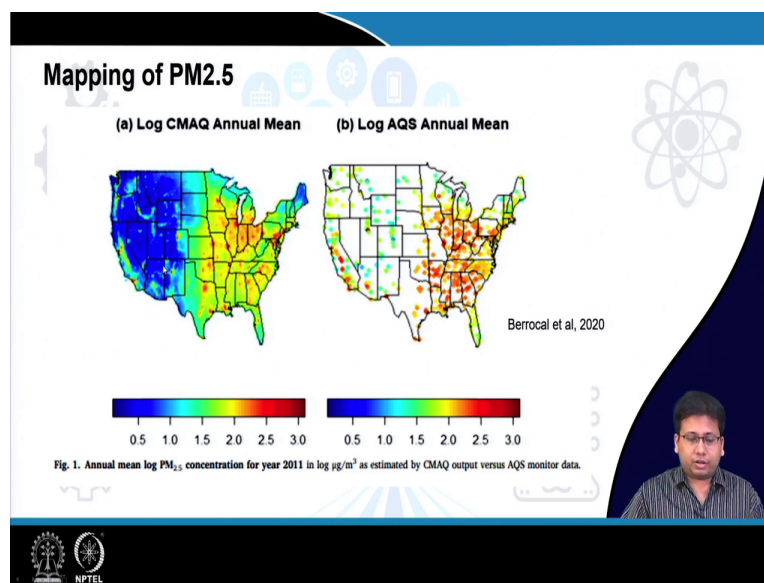
So, in this paper they have done it and identified certain specific covariates, and then there can be some additional covariates also. And, finally, they have the output from CMAQ the process model which itself may be used as a covariate I mean it need not be, but it can be also used.

Next thing the next is all these things can be fed into different approaches. So, first of all there are different machine learning methods such as I said some SVR random forest etcetera which they have considered. Apart from that they also have IDW which is Inverse Distance Weighting which it is basically some kind of a spatial linear spatial interpolation technique then there is kriging statistical downscaling etcetera.

Now, the thing is when we are considering the like the vanilla machine learning algorithms like random forest etcetera we do not consider the spatial properties. We at any given point we try to make a point wise prediction of the like of the variable of interest; in this case the air quality, but we do not take the spatial dependence while IDW, kriging etcetera they actually take the spatial the spatial dependence into account.

So, like this is the universal kriging algorithm which they have considered. So, like here you can take a look at it.

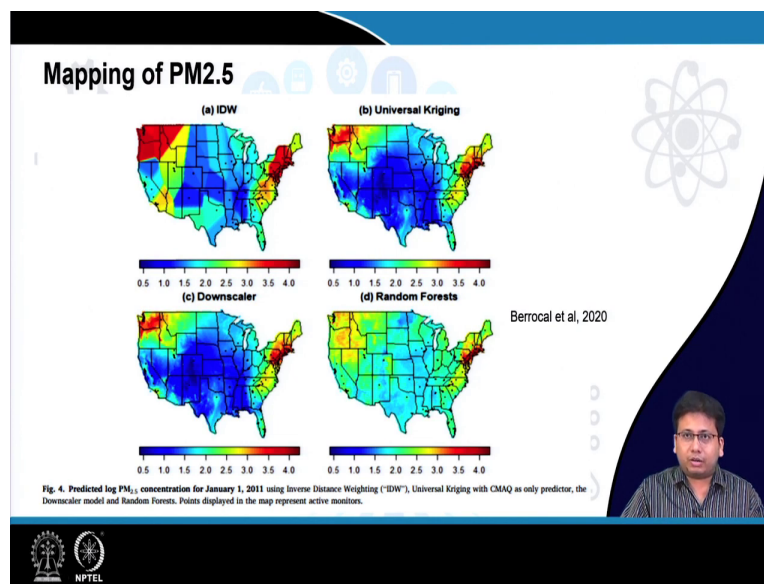
(Refer Slide Time: 32:00)



So, now like this basically shows the map of this $PM_{2.5}$. On this side you can see what is obtained from the CMAQ and like these are the observations. So, like you can see this point wise observations. These the points which you can see these are the stations.

As you can see that there are not only are they sparse, but they are unevenly distributed in the like in this east coast as well as the west coast there is a dense region, but the in the vast parts of Central America there is there are very few of such observation stations. Now, this like, but if you somehow using the CMAQ process model like if you are able to do it you can see this kind of like this kind of a map.

(Refer Slide Time: 32:46)



And, now like so, here is we are comparing the maps obtained from different approaches. This is IDW – Inverse Distance Weighting, this is universal kriging, this is downscaler another like another algorithm and this is random forest which is a machine learning. And, as you can see in all the different cases we get significantly different maps. The like for example, in case of random forest we see like except for a few locations it is like almost everything is seems roughly the same.

In some of the like in this case downscaler like here most of the region seems to be having very low while in this case in case of IDW we see significant amount of spatial variations, in this universal kriging also we see that. And, it actually turns out that like when they plot the results it turns out that like the best results are obtained from the IDW universal kriging etcetera and not from the machine learning algorithms.

The in fact, that is because the machine learning model the models random forest etcetera do not in the in their original form they do not take into account the spatial dependence which is actually taken into account by kriging and so on. Further it is shown that if these CMAQ outputs if these are used as another of the predictors that actually helps in the prediction.

That is, rather than depending purely on the these meteorological covariates if these CMAQ outputs are also taken into account while making the predictions that actually helps the prediction improves the accuracy.

(Refer Slide Time: 34:34)




REFERENCES

- Berrocal VJ, Guan Y, Muyskens A, Wang H, Reich BJ, Mulholland JA, Chang HH. A comparison of statistical and machine learning methods for creating national daily maps of ambient PM_{2.5} concentration. *Atmospheric Environment*. 2020 Feb 1;222:117130.
- Kathuria D, Mohanty BP, Katzfuss M. A nonstationary geostatistical framework for soil moisture prediction in the presence of surface heterogeneity. *Water Resources Research*. 2019 Jan;55(1):729-53.



(Refer Slide Time: 34:36)



KEY POINTS

- Soil Moisture exhibits considerable nonstationary behavior at subgrid scales, and stationary geostatistical methods fail to capture this behavior
- Surface controls like vegetation and rainfall can be used to construct covariance functions that account for nonstationarity
- Statistical methods like Kriging can outperform ML algorithms by better utilizing spatial information

So, basically the point to be taken home here is that are as follows. So, first of all the soil moisture exhibits in the first paper so, we found that soil moisture exhibits considerable non stationary behaviour at subgrid scales and stationary geostatistical methods fail to capture this. Surface control like vegetation, soil texture and covariates like rainfall, they can be used to construct covariance functions that account for the non stationarity.

And, finally, statistical methods like kriging, they can outperform vanilla machine learning algorithms by utilizing spatial information. So, like this is not a criticism of machine learning algorithms here it is just to say that machine learning algorithms in their original form may not be suitable like very successful in this in case of this spatiotemporal data unless they take into account the spatial information. So, that is why it is necessary to tweak them in that in that particular way.

So, with that I come to the end of this lecture and we will continue this discussion of various using various machine learning methods to answer questions in Earth Sciences in the coming lecture.

So, till then, bye.

