

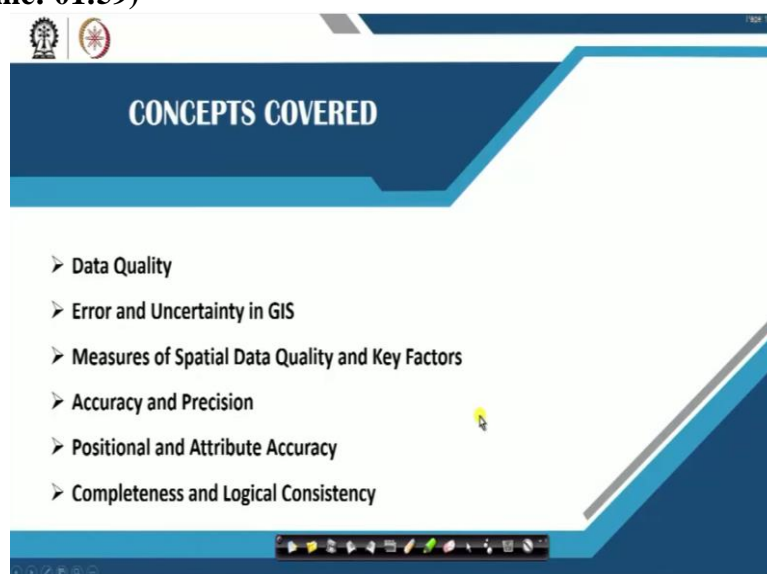
Lecture 21
Data Quality and Measures

Hello and Namaste. Welcome back to the course on geographic information system. So in the last week, we have learned about data models, what are different kinds of data models? How do we handle data? And now, the third week, we looked at how the data is generated. Right. So the next week, we have looked at how the data once the data is generated, how do we handle it in terms of data model?

Now, whatever the data that has been generated, so you need to understand the data quality. So to understand this data quality, in today's this week, specifically, we would look at what are the different data measures, data quality, data assessment techniques, errors that may have. So these are actually the overall errors where most of us would have it. But I am not going to some very, very detailed, specific details, where in which actually say that is this may be an error, but I am giving you an entire list of errors and quality checks that the entire database or the entire population who may be handling GIS may have.

So, this is not very specific. This is for the entire population. But only thing is that if you can understand that. What are the necessities, what maybe there are, then it would be really good enough for you to understand how that has become and how do you mitigate it.

(Refer Slide Time: 01:59)



Okay, so in today's class, in today's class, let us look at our data quality, error and uncertainty in data in GIS. There are certain uncertainties associated with GIS, how the data is stored, how the data is manipulated, so we look at it and then measures of special data quality and key factors. So we have to also look at how do we measure a data quality. So if there are certain data measures are the factors that are governing this quality, then we have to look at those, then accuracy and precision.

This becomes extremely important when you are actually putting out this in a public domain. If you do not say what how much accurate is your data, and it may not be the users may not be able to adopt it for the real scenario when they are trying to look at your model. But if you have to be precise, you have to be accurate. So, how much accurate are you. We are not all the data that I said in my previous classes can replicate 100% of the real world, but if it can replicate a certain percentage of the real word, then it is assumed to be accurate to that extent.

So, what are those measures? We look at it what are the present factors also we look in this particular slides, then we look at what is positional accuracy and attribute accuracy. Both of these are extremely important in terms of when you are looking at the database then finally, completeness and logical consistency is your car your database or GIS data that you have already put in the database complete or has certain consistency has some logic error if the logic is missing and the logical consistency is missing, then your database may not provide us the best results.

(Refer Slide Time: 03:44)

Data Quality

- The meaning of 'quality' depends on the context in which it is applied
- Quality is more difficult to define for geospatial data
- Unlike manufactured products, geospatial data do not have physical characteristics that allow quality to be easily assessed
- Quality is thus a function of intangible properties such as 'completeness' and 'consistency'

Source: aibook.in

NPTEL

So, how do we look at it will also go through in this particular set of lectures. Now, first when we look at data quality, the first thing when you always see the term quality. So, one thing is that when you go and buy grocery, so, you try to look at the quality maybe in terms of price, maybe in terms of certain brands, maybe insight in terms of certain size of that particular product or size and shape or the color of that particular product when you are normally buying in a grocery store.

But in GIS, it is extremely difficult in terms of understanding data quality ok. So, the data quality is more contextual in terms of having a data. Now, when you are looking at manufactured products, I said now geospatial data do not have any physical characteristics like your color, shape or maybe a brand. So, there is no physical characteristic that allow easier the quality to be easily assessed. Only thing that you have is the meta data, so meta data, you see the data about the entire data.

So when you are looking at this quality, so how do you how do we actually look at a data quality should we look at in terms of just a measure on are we looking at in terms of its different data sources, data measures, that is what we mean by quality here, we nowhere there is a comprehensive list of the things that has to be satisfied in order to have a good data quality, as far as GIS is concerned.

Quality is just a function of an intangible property such as completeness and consistency. So, how complete is your data how consistent is your data. If someone had generated the same for example, a government agency generates the same data for the next year what you have generated this year. So, when you look at it, the because of a government agency may have done and the entire field they may have the entire field knowledge and have done the entire fieldwork and in this year.

So, they would have got a certain amount of data that is completely true on ground. So, the same quality of data that you have published, if that there is a consistency for example, where there is no change if the consistent is there, and where there is change. If that is against consistent, then it means to say that your data is actually good. So, similarly it depends on what is completeness, what is consistency. So, we look at what are different measures of completeness and consistency also.

(Refer Slide Time: 06:32)

Error and Uncertainty in GIS

- One of the major problems currently existing within GIS is the characteristic of accuracy of digital geospatial data
- Error propagation in GIS is mainly due to the process of integration of several data sets, at different scales (spatial and temporal resolution) and quality
- The ease with which geospatial data in a GIS can be used at multi-scale signifies the importance of data quality information

Now, the thing that always that creeps into our mind when you are looking at any data is error and uncertainty. So, always any data whatever irrespective of whatever the tool software etc you are using has certain error has certain uncertainty. So, that is one of the very major problems that is currently existing within GIS and this characteristic is a characteristic of accuracy of a digital geospatial data.

So, that is why we will say this data is so much percent accurate it cannot never GIS data can we make a real world it cannot be 100 % accurate it cannot be 100 % true. It cannot really make a real world because real world is a continuous phenomena and you can see they have their own representations discrete representations. But when we are looking at here we put it into real data in terms of converting it from a discrete continuous phenomena to only a discrete phenomena as we may be line, polygon etc.

So, hence it is not easy to mimic a real world, but it is also that you cannot say you have 100 accurate data because it absolutely does not represent a real world. Then error propagation in GIS is mainly due to process of integration of several data sets. For example, let us say I am I may have shown you one of the data sets where I have urban data I have a radius data I have a population data I have a CDP data, I will have engrossment data or all of this data that has been together so these are from different sources.

If the If it is a single user who is following a single way of representation of these are different layers, then whatever the data that has been collected, may be significantly in a standardized form that he is following, but let us say that there are different users who are

generating this data and this and one user is making use of all these different sources in order to propagate the entire model and build a model.

So in that case, there may be several errors. Errors are different and standardization techniques that people have used in our so that becomes a very, very important aspect of a digital GIS. So, to control that we use certain measures so that the error is not propagated to the entire data set. So the ease with the geospatial data in the GIS can be used at a multi-scale signifies the importance of data quality information.

So, whatever the data that you have in this particular database and you can integrate any number of databases and also you can use it a multi scale information and with the same data, so, if you are trying to do that, then the information should be accurate to a very large extent.

(Refer Slide Time: 09:46)

The slide is titled "Errors in GIS" and lists several possible sources of errors. On the right side, there is an icon of a map with a red 'X' over it, indicating an error. The slide also features a small video inset of a person in the bottom right corner and a navigation bar at the bottom.

Errors in GIS

- Possible sources of errors in GIS include:
 - Wrongly assigning the coordinate systems (often done by users)
 - Mislabeling of areas on thematic maps
 - Misplacement of horizontal (positional) boundaries
 - Mismatch in spatial and attribute data
 - Human error in digitizing
 - Classification error
 - Human bias

What may be the possible errors in GIS. So, there are a huge number of possible errors that GIS user can have, but I have listed a few of them, which are major errors. So I have not gone into every detail of this maybe if you have polygon like this or polygon closing so I am not going into all those details but I am signifying what may be the major errors, which lead to your error propagation in the entire model.

Ok. So now when you are looking at errors in GIS, you may have wrongly assigned a coordinate system. So, a person has generated 1 GIS layer other person has generated another GIS layer. So both of them have a different coordinate system. As I previously said in

my previous class when I am explaining about the coordinate system both may not sit only on each other are set above each other, ok.

So, in order to do that, if you have to do a top down analysis or bottom up analysis, you need to have provided a good compatible or maybe the same coordinate system. So that it is easily so comparable. Then mislabelling of areas in a thematic maps for example, our generated thematic map I have 4 categories, let us say I would say let us say I have 3 categories, which is urban, vegetation and water.

So 0 first classes urban second class is vegetation, third class is water. What now one of my colleagues develop another map wherein he or she has first class is vegetation, second is urban and third is water. And there is another colleague who develops another map. So, he or she has first class is water second class is urban third class is vegetation. Now, you have 3 different ways of representation of thematic maps.

When you are trying to visualize the entire data set, you may not be able to get the exact information or the information that you would get is actually inaccurate or wrong, so if this mislabelling is done is not is not done. So most of the times most of the error comes through mislabelling, please look at it and have the same labels for the entire data set that you are actually using.

I am not saying that every map should have the same data. But if you are comparing, for example, if I am looking at urbanization process, anything that is concerned to urban growth, so I label it in a certain form in the first map, the same thing goes to the 100th map or the end map. So, if that convention is followed most of the errors are removed. Now, the next one is misplacement of horizontal boundaries, the positional boundaries many of the times.

So, that this is most inaccurate way because many of the times people have different sources of extraction of boundaries. So, I have already said you can extract boundaries through Google map or you can extract boundaries through Google earth or you can extract boundaries through Survey of India toposheets. So, people have different ways of extraction of boundaries.

So, when you are extracting the boundary, you have to be extremely again careful with its meta data, look at the metadata. If all of the maps are generated with the same detailed boundary description, then it will be fine enough. Then mismatch in spatial and attribute data this is another very important wherein the mismatch happens with either a spatial data or the attribute data many times it has spatial data has been attribute data.

Human error and digitizing: So, we have discussed about these errors. So, when you are digitizing certain polygons maybe a lake maybe a tree maybe a building, so, there may be a certain errors that creep in because of your digitization. So that errors may also lead to errors in GIS. That also tells you that if you give a positional information that shows what is the error and it may not show where is the error, but it shows how much accurate is your data, then the classification error.

For example, when you are doing a large roster land use land cover classification, when you are using land use cloud classification, the classification error is a very important point where you are trying to understand how accurate is your data, whatever the data that you are putted in a public domain has to be extremely accurate. It has to be accurate. Then, your classification to be very good.

In order that that this classification is very good, the error has to be minimal, which means the accuracy of the land use that you have generated should be validated in a proper way and more or more importantly that signatures that you have used must represent the entire heterogeneous classes. So, we will look at all of this will look at some of at least some of these when we are looking at it in the practical class, then human bias.

So, there are certain biases for example, ah if I am from, let us say, from a particular city, so, my bias is towards something, or if I am looking at only urban phenomena and not the water phenomena, so, my bias is towards urban phenomena. So, the more detailed information is towards urban phenomena whereas the water is neglected. So, that kind of bias creep in which is actually a major source of error. So, all of these put forth from scenarios.

So, in order to avoid or inform the user that how good is your data not any of the data and that is 100 % accurate. So, we have to inform the user what is accuracy of your map.

(Refer Slide Time: 15:34)

The slide is titled "Measures of Spatial Data Quality" and features a list of measures. The background includes a stylized tree diagram with various icons and a small video inset of a man in a checkered shirt. The NPTEL logo is visible in the bottom left corner.

- The most important measures of data quality are,
 - Data accuracy
 1. Positional data accuracy (Absolute and relative)
 2. Attribute data accuracy
 - Data consistency
 - Data completeness
 - Data timeliness
- Other concepts used are
 - Lineage
 - Accessibility

So, when you are looking at any of the measures if someone asked me you have so many errors. So, how do you measure an accuracy of your spatial data since you said it as a subjective term, there vary a measures of looking at spatial data accuracy, for example, when I say data accuracy, it may be a positional accuracy it may be attribute data accuracy. So, when I look at positional accuracy, it is absolute and relative, then we have data consistency.

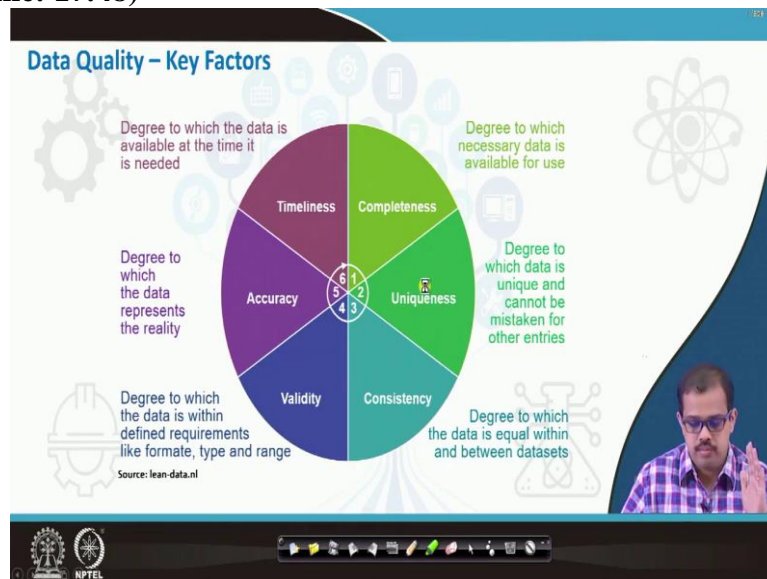
So, data consistency is extremely important how we will measure will also look at that data completeness is one measure of data quality than data timeliness. So, if you have if you are generating a regular data. So, the regular intervals of time is extremely important in order to understand what the data set actually combines. Then there maybe some more concepts like lineage and accessibility, when I say lineage, it is how the data has progress from its birth or from where it has been born and killed the usage, how it has progressed over time.

How it has been improved, degraded etc, over a period of time. So that is what is called lineage. Then you have accessibility. Accessibility is another factor when in today's context. Also though, we have a huge set of data that has been put out some very important crucial data about maybe a land area maybe about the statistics of a city, etc. Many of them are getting close domains. Though you can get so many details in Google look at Google database look at extract information from toposheets through virtual globes etc.

But then also you have certain limitations on accessibility of this data which is extremely which may be extremely accurate. So, such also such things also looking at it we can say

what kind of data quality it has if it is quite accessible, we can compare the data quality and understand how good the quality of generated map is there.

(Refer Slide Time: 17:48)



Now, I say fire as I said so, when we look at it layer, we have to look at it in 6 different ways. The first way as I said there is completeness when I say completeness it is degree to which are the necessary data is available for use. Ok, next uniqueness. So, it means that it has a degree to which data is unique and cannot be mistaken to any other entities. Then you have consistency degree to which the data is equal within or in between 2 datasets.

So, they are this is what is consistency then other one is validity. The degree to which the data is within the defined requirements like formula of our format type and range. So, this is also extremely important and when you are looking at accuracy degree to which the data represents the reality. So, that is what we are trying to look at them or look mimic the real world. So, this is what is extremely important in terms of making the real world and last but not the least it is timeliness the degree to which the data is available at the time it is needed and the amount of data that is available over a time bound period.

So, that is also extremely important in terms of data quality. So, please remember these are the 6 key factors when you are looking at that GIS data when you are looking at revolution of data. So, look at all of these aspects.

(Refer Slide Time: 19:19)

Aspects Considered for Data Accuracy

- Aspects of data acquisition considered for data accuracy are
 - Needs
 - Costs
 - Accessibility
 - Time frame
- Quality can be defined as one or more characteristics of geospatial data that describe the extent to which it is fit for use

The slide features a background with a stylized tree of icons representing various data-related concepts. A presenter is visible in a small video window on the right side of the slide. The NPTEL logo is located in the bottom left corner.

So, once we have understood what is data accuracy will look at what are the aspects that are considered for data accuracy. For example, the first thing that actually starts with when you are measuring a data accuracy is the aspects of data acquisition. So, how the data is acquired, what are the needs of acquiring that particular data, why the data has been acquired. Then the cost that is involved in acquiring the data.

Accessibility and the time frame other than this the method at which it has been acquired also matters, though I am not listed here method is also a very important factor in terms of acquisition of a data. Then we are when we are looking at quality, I say informed to be here before quality can be defined as one or more characteristics of us geospatial data that described the extent to which it is for us. That is it.

Okay. So, it may even if let us say that there is no data available at all, you have a bad quality data then also that data is good for us. But if you have huge amount of data and the same source in the same domain, and if you have a bad quality data and if you have a good quality data, then I will obviously you will go for in a good quality data in order to see how good that particular data is. So, that is; or utilize that data for to extract some information out of it.

(Refer Slide Time: 20:54)

Accuracy and Measurement of Accuracy

- Accuracy is the degree of correspondence between data and the real world, **Fundamentally controlled by the quality of the input**
- Measurement of accuracy
 - Standard deviation is considered as a measure of accuracy
 - The Standard deviation (SD) is the measure of how close a measurement/observation lies to the value expected

$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Now, when you are looking at accuracy and measurement of accuracy, so when we define accuracy, so you should be extremely careful, it is the degree of correspondence between data and the real world. What I mean to say is the data model and the real world model, fundamentally controlled by so here I am very specific. It has fundamentally controlled please be careful here fundamentally controlled by the quality of input, if your quality of input is wrong or bad, then your entire accuracy is bad.

Okay. So, many of people when you are when they are doing the land use classification they come back to us and say sir, we have we could not achieve a good accuracy of Land classification exactly, because the amount of field investigation that is involved even looking at visually the scenarios that I see in the real world and trying to propagate into the data model, we are looking at the real world data collecting the data.

So, all of these would be missing. So, that is where the quality of input when I said quality of input, you have to have certain field measurements in order to see that your data set whatever you have collected, does not have any bias. So, once you have collected certain data from the ground and have put into the database along with that, you have certain data that you input based on your perception based on various degree of perceptions.

So, all of this stuff together will give you a better accurate data rather than using just a biased data. So, always remember that it is your input that actually matters, when you are looking at the quality then when you are looking at the measurement of accuracy, so, now, there is

standard deviation is considered to be the measure of accuracy basically, first thing that you do is that you populate the entire database.

Then, once you have populated it, you look at what is a mean for example, following at the land use classification look at the mean then you look at the standard deviation of that particular and data. So, you know, what are the different out layers in that particular data set which can be either removed or which can be maintained in such a way that it provides some critical inputs during your classification most of the time these out layers are removed. Ok. So that is about accuracy and measurement.

(Refer Slide Time: 23:38)

Error Ellipsoid

- In surveying, error ellipse (2-dimensional) and error ellipsoid (3-dimensional) are utilized to visualize the uncertainty of points
- Error ellipse has semi-major and semi-minor axes equal to standard deviation in each direction
- The probability of point lying outside the standard error ellipse is much larger, Therefore ellipsoids are used for better results
- The likelihood of observations/measurements lying within given boundary is known as confidence and is expressed as percentage

The diagram shows a 2D coordinate system with X and Y axes. A 'standard error rectangle' is drawn with dimensions $+S_x$ and $-S_x$ along the X-axis and $+S_y$ and $-S_y$ along the Y-axis. Inside this rectangle, a 'standard error ellipse' is drawn, centered at the origin. The ellipse's major axis is labeled 'major axis' and its minor axis is labeled 'minor axis'. An 'adjusted point' is marked at the center of the ellipse. The source is cited as 'Autodesk'.

So, the next thing that we would look at is the error ellipsoid. So, this is another basic thing that people consider to be a mistake. So, how to said it actually happen for example, when you are surveying error ellipse and error ellipsoid are utilized to visualize the uncertainty points, when I say error ellipse it is a 2 dimensional way when I say ellipsoid, the ellipse rotated about its central axis or the primary axis.

So, this is nothing but a 3 dimensional figure. So, when you are looking at error ellipse. So, for example, when you look at this here, so, this particular thing here is nothing but you are minor axis. Whatever you see here is a minor axis when you look at this this is a minor axis and this is your major axis. So, this is your ellipse that we have considered. Now, with this if this is the ellipse that we have considered.

Now, let us draw a box that is actually connecting the 4 different points that are one is the further away and other one is a closer so 4 different points as 2 pairs. So, once you have done this the probability of any of the point that is lying outside the standard ellipse is much larger. If these ellipsoid are used for better results, so, when you are looking at the probability of that particular point if the probability of point here then that particular point as in a correct shape.

Whereas if it is a probability of a particular point outside this particular ellipsoid, then you probably have to look at what kind of data sources you have considered. So, when you are looking at this particular observation, the likelihood of observation or measurements lying within the given boundary is actually known as confidence, higher the number of confidence, which is expressed in percentage higher amount of confidence, better is your data, lesser the amount of confidence, less is your data.

If you have huge number of data points that is lying inside this particular area, then you have your data is accurate enough to consider for the analysis. Yes, if there are some data points which are outside these are those erroneous pixels which may not be actually consider erroneous data points. So, looking at what is the amount of proportion that gives you exactly what is the error ellipsoid. So, that is why we draw a error ellipsoid and look at what is the proportion of points that is getting into this left side and what is the proportion of point that is not in the error ellipsoid. So, this how you measure your error as an factor.

(Refer Slide Time: 26:38)

The slide is titled "Random and Systematic Error" and contains the following content:

- Deviations observed may be random/systematic
- The systematic errors are computed according to the following equation
- The systematic component may not have much importance if the data is originating from the same source

$$m = \sum_{j=1}^n (X_j - \mu) / n$$

The slide also features a diagram of a square with a diagonal line and a pink ellipse drawn over it. A small inset video shows a man in a plaid shirt speaking.

Next one another thing is the random and the systematic error deviations are normally it is random, but many times it is systematic, but 90% of the time it is random. So, these random

errors can be corrected through various measures, I would not go into details of these measures, because these are not these are this has to be looked at when you have the entire data set. But when you look when you are looking at the systematic data are ah computed using various measures for example, ah now using this particular equation,

okay. So, that gives you the how the systematic error has been propagated or through the same source it is originating most of the systematic error originated from the same source keep that in mind.

(Refer Slide Time: 27:27)

The slide is titled "Systematic error" in blue text. It features a background with various icons related to data and science, including gears, a tree, a hard hat, and a beaker. A video inset in the bottom right corner shows a man with glasses and a mustache, wearing a purple and white checkered shirt, speaking. The slide contains the following bulleted text:

- Serious errors occur in data and must and have to be removed, the definition of what causes gross blunders
- Redundancy of observation is used to remove blunders, outliers which are probable blunders are removed by statistical testing
- Errors may also appear in data processing

At the bottom of the slide, there is a navigation bar with several icons and the NPTEL logo on the left.

So, that is how you look at the random error and systematic error, but when you are looking at systematic error see it is the serious errors that occur in the data it may be due to malfunction of a particular sensor it may be because of the day in the error in the type of sampling that you are trying to do. So, these are the errors that have to be always removed. So, if this errors are not removed, then you are probably having the data set your data may have certain issues when you are processing the entire data.

So, which means to say that you may give wrong outputs or wrong decisions when you are actually and trying to understand the top down approach or the bottom up approach. If you have certain issues like this redundancy of observation is used to remove this blunder. So redundant observations many number of observations can be used to remove these blunder outliers which are probable issues are to be removed by statistical testing.

Errors may also apply appear in the data processing part. So, that kind of errors as well I mean it is user specific it can I cannot be looked at in terms of a theoretical aspect, but data processing errors also may creep in when you are looking at the systematic error point of view.

(Refer Slide Time:28:45)

Precision

- It is the exactness of measurement or description
- It also expresses repeatability of the measurements
- The “size” of the “smallest” feature which can be displayed, recognized, or described
- Can apply to space, time (e.g. daily versus annual), or attribute for raster data, it is the size of the pixel (resolution)

Accurate Precise Not Accurate Precise Accurate Not Precise Not Accurate Not Precise

The slide contains four target diagrams. The first target has a single red dot in the center bullseye, labeled 'Accurate Precise'. The second target has multiple red dots scattered around the center, labeled 'Not Accurate Precise'. The third target has multiple red dots clustered together but far from the center, labeled 'Accurate Not Precise'. The fourth target has multiple red dots scattered far from the center, labeled 'Not Accurate Not Precise'. A small video inset in the bottom right corner shows a man with glasses and a plaid shirt.

Next is precision, it is exactness of the measurement of a description. For example, I have different circles that I have put in here, when you are trying to actually provide a point that is actually fitting the first particular circle of center circle than it is accurate and precise. At the same point set somewhere else that is inaccurate for example, in the all of these 3 data sets. For example, when you see in the second case, this is neither accurate nor precise.

But when you look at the in the third case, this particular thing is accurate enough, it is closer to whatever is the particular model, but it is not precise, a certain number of a certain amount of accuracy it may be 80, 70, 60% accurate, but it is not precise, but the first figure is precise then the last one is neither accurate nor precise. So, you have to look at what kind of data that you are actually handling and how this data is either accurate or precise.

So, it also expresses the repeatability of measurements the size of the smallest feature which can be displayed, recognized or described that is very important when you are looking at precision. It can apply to space, time, for example, daily versus annual or attribute for raster data that is the size of also a pixel for example resolution, so, We have to look at these aspects when you are looking so whether it is accurate and precise whether it is accurate but

not precise whether it is precise accurate or not accurate or both it is not accurate and not precise. So look at all of these possibilities when you are testing your GPS data.

(Refer Slide Time: 30:37)

Positional Accuracy

- Positional Accuracy (sometimes called Quantitative accuracy)
- Defined as the closeness of locational information (usually coordinates) to the true position
 - Spatial
 - Horizontal accuracy: distance from true location
 - Vertical accuracy: difference from true height
 - Temporal
 - Difference from actual time and/or date

The slide includes a map of a region in India with a blue highlighted area and a video inset of a presenter in a purple and white checkered shirt.

In the last 2 things that we have to look at when you are looking at accuracy is positional accuracy. When you are looking at positional accuracy It is nothing but a computed quantitative accuracy. So it is defined as a closeness of locational information to true position. So when you looking at this personal accuracy are 2 types, one is spatial other one is temporal. When I say spatial, it is it again is divided into 2 types one is horizontal accuracy, another one is the vertical accuracy.

When you look at the horizontal accuracy, it is a distance from the true location that is displaced horizontally, whereas vertically, it is the difference from the true height. That displays an amount of displacement that it has in the true height, so difference from the true height and the displacement that it has from the true height. That is nothing but it is vertical accuracy. So we look at both one in the 2d model, other one is a 3d model. And when you are looking at temporal, it is a difference from the actual time or date.

Okay, actual time of your sample collection, maybe today at 10am. But your data that has been delivered by the land side, maybe at 10:30pm where the sampling time, so the half an hour difference is your virginal accuracy of that particular point. So there may be certain things that are dynamic would have changed over an hour, 30 minutes. So when you are looking at when you are looking at your data, you should have you should think about that kind of change that may have happened over a period of time.

(Refer Slide Time: 32:19)

Attribute Accuracy

- Attribute Accuracy or Consistency- the validity concept in experimental design
- Defined as the closeness of attribute values to their true value
 - A feature is what the GIS/map purports it to be
 - A rail line is a rail line, and not a road

The slide also features a map in the top right corner showing a 'Rail line' and a 'Road' in a residential area. The presenter is a man with glasses and a mustache, wearing a checkered shirt, pointing at the slide.

Then the next thing is attribute accuracy. Attribute accuracy or consistency is the validity concept in experimental design, how valid is the concept in your experimental design. Then defined as closeness of attribute values to their true values how close is attribute value to its true value that is a feature in what the GIS maps is supposed to be or for example, in this particular line a railway line is the railway line itself and not a road. So, that is what is so, you cannot assign a railway line as attribute value to be as a road then it is a mistake.

So, if that kind of accuracy are not verified, then such a attribute accuracy may not pose a great challenge in terms of if you have a good knowledge about the entire space and the entire real world then attribute accuracy may not be a big challenge, but in terms of its huge database having huge amount of data, then attribute accuracy poses a great challenge because it has to be populated in a proper way.

(Refer Slide Time: 33:37)

Completeness

- It is the reliability concept from experimental design
- Are all instances of a feature the GIS/map claims to include, in fact, there?
- Partially a function of the criteria for including features: **when does a road become a highway?**
- Simply to put, how much data is missing?
- Completeness refers to and absence of features, their attributes and relationships of spatial data in comparing data defined in the data model or the representation of the real world

NPTEL

Though, as I said the last part of it is whether the data in some database is complete, it is the reliability concept for an experimental design for example, are all the instances of a feature in a GPS or a map or a store value in a map claims to include in fact that is there. So, that is what is completeness, whether that is in the real world model.

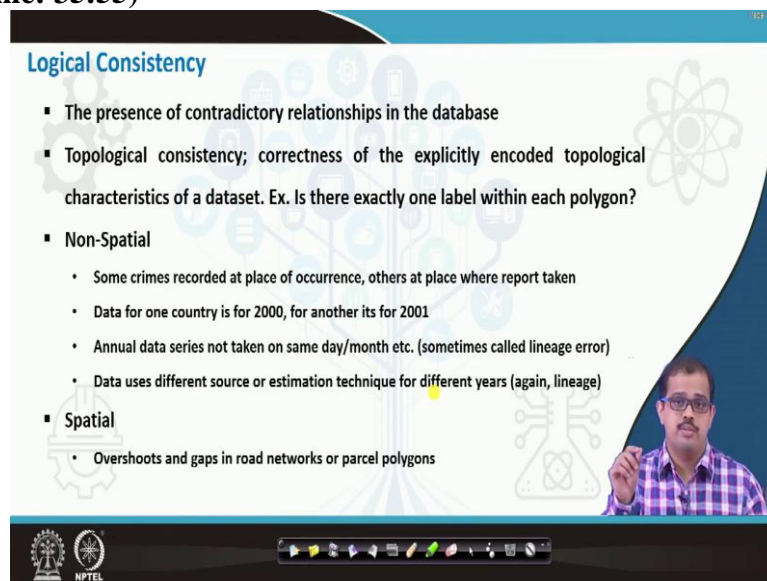
So, if everything is as is correct, then your data is kind of complete, then partially a function of criteria for including features that is when does the road become a highway, what kind of information that is actually complete then whatever said and done, if you want to measure completeness, look at the data that is missing, if you can satisfy that criteria of what data is missing, then you are most of your database or the data model is complete.

Then the complete is also refers to and the absence of features when I say features, these had the data model events in the data model or entities in the data model, their attribute information whatever the attributes are stored, their relationships with the spatial data, what are the relationship that has been built, if there are certain relationships that are cross cutting and are not there, then you have to build upon that in are in comparing data that is defined in the entire data model.

Then if you have understood the entire space, that is the spatial aspect, then the database along with the data and its relationship or the entity and its relationship, then your data is complete. If the entire thing is not validated, and it is not correct to the real world, then you

have incomplete data. When you have incomplete data, it means that your database is most inaccurate.

(Refer Slide Time: 35:55)



The slide, titled "Logical Consistency", lists several database issues:

- The presence of contradictory relationships in the database
- Topological consistency; correctness of the explicitly encoded topological characteristics of a dataset. Ex. Is there exactly one label within each polygon?
- Non-Spatial
 - Some crimes recorded at place of occurrence, others at place where report taken
 - Data for one country is for 2000, for another its for 2001
 - Annual data series not taken on same day/month etc. (sometimes called lineage error)
 - Data uses different source or estimation technique for different years (again, lineage)
- Spatial
 - Overshoots and gaps in road networks or parcel polygons

The slide also features a presenter in the bottom right corner and a navigation bar at the bottom.

So please look at the suspect. So this is where most of them have an issue when they are developing an entire model. So, they are not looking at the entire boundary. First, what is the aspects that you consider in your model and what is the aspects that are not they have entire set of aspects that they consider. And finally, they will not be able to find out certain entities that are the information about certain entities at certain places, which we needed for to complete the entire database. So, look at the completeness of the database.

The final thing is logical consistency. So, the presence of a contradictory relationship in the database, the topological consistency that is there in the entire data and the correctness of explicitly encoded topological characteristics of a data set. So, this is extremely important when you are building the data set. So, when you are looking at the topological data set, please see that how consistent this topological data set you cannot have a topological data set with a sudden uneven way of representation.

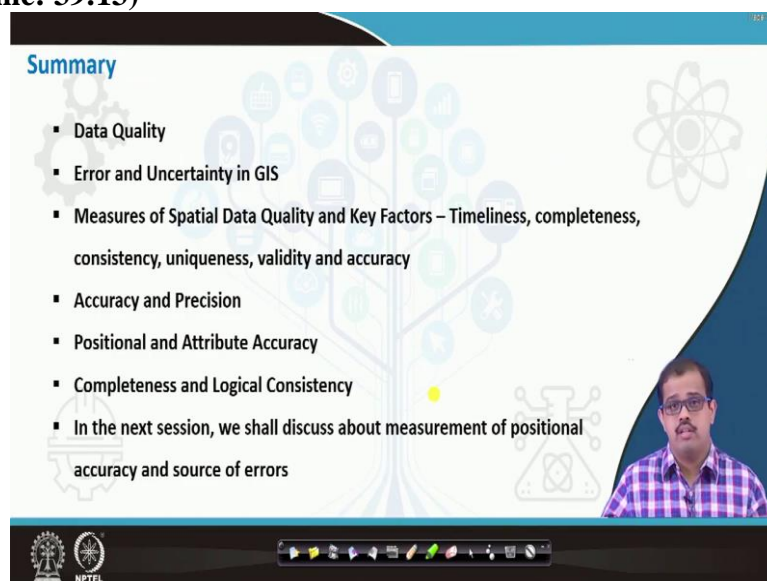
So, please look at if there are certain things like that corrected, edit the entire data, edit the data and correct such issues. So, when you are looking at this there is non spatial and spatial when you are looking at non spatial, for example, just to give an example some crimes that are recorded at a place of occurrence and others Where report is taken. So, sometimes a crime happens and immediately it is recorded by the police are some of the crimes that are reported in the police station.

So, these are non spatial entities, but it should have a logical consistency data for 1 country is 2000 and the data for the other countries for 2001. So, which again does not have a logical meaning in terms of consistency. So, you cannot compare both of these both the country should be the data should be comparable for 2000 or 2001 whichever you have then annual data series not taken on the same day or month.

That is also extremely I said the time the month the minute is extremely important when you are actually comparing to data set especially if you have remote sensing data, the time, the day etc. plays a very, very important role in terms of what kind of data you have and how you are analyzing the data and comparing the data. Then data uses from the different sources are estimation techniques for different years. So, lineage is very important aspect then look at how this data is involved also is important.

And when you look at spatial characteristics overshoots that we have seen, there are a lot of errors that when we are actually digitizing, overshoots all these have different errors. So look at those errors that may have creep into the data, so, correct those errors. So to have a logical consistency otherwise if we are trying to find out the area of polygon in the entire data set that you have already generated, you may not be able to find out that certain polygons which have these errors, whether it is overshoot, or under or undershoot or having gaps. So that will give you a lot of errors when you are trying to calculate the area.

(Refer Slide Time: 39:13)



Summary

- Data Quality
- Error and Uncertainty in GIS
- Measures of Spatial Data Quality and Key Factors – Timeliness, completeness, consistency, uniqueness, validity and accuracy
- Accuracy and Precision
- Positional and Attribute Accuracy
- Completeness and Logical Consistency
- In the next session, we shall discuss about measurement of positional accuracy and source of errors

NPTEL

So these are certain data quality checks the errors that you may encounter in a larger issue when you are actually looking at, but when you start using that GIS data and start populating your database, I am sure you will find out huge number of errors that may creep in. So, every for every error they send, I mean, troubleshooting that you need to do and maintain the entire database in a proper way.

So, if there are huge numbers of sources, gets it a common platform, see that your data is interoperable and use it in the way that it has to be applied. So, keep this in mind look at the standards look at what are the different ways of handling the data and then use that particular data otherwise you are going to miss with the entire data set. To summarize this particular class, we looked at what is the data quality, what do you what do we mean by data quality?

We understood that data quality is not a measurement for a GIS in terms of physical aspect, but it is a measurement in terms of how the data is represented, then we looked at errors and uncertainties in the data there are certain errors there are certain uncertainties it is the way we look at we also look at how do we capture these errors in my probably next lectures and then accuracy and precision.

So, not all data should be both accurate and precise or it not both data not all data will be both not accurate not precise, but it may be accurate, but not precise. It may be precise, but not accurate. So, all of these things are possible. Then you have positional and attribute accuracy, what is positional accuracy, how the positional accuracy is extremely important and how it is quantitative.

Whereas attribute accuracy is the quality of the attributes that you have, where you have neither have not connected with relationships or entities are not represented properly. So all of these are very important in terms of looking at data and data quality and very, very important thing is that it is complete and consistent. So, if the data is complete and consistent, then then this data has certain value when you are applying it to your the real world.

So, probably in the next session, we will look at the measurement of positional accuracy, the sources of errors, and how do you overcome this errors in next 2, 3 lectures that I would handle after this class. Till then, let us meet in the next class. Thank you very much. Have a nice day.