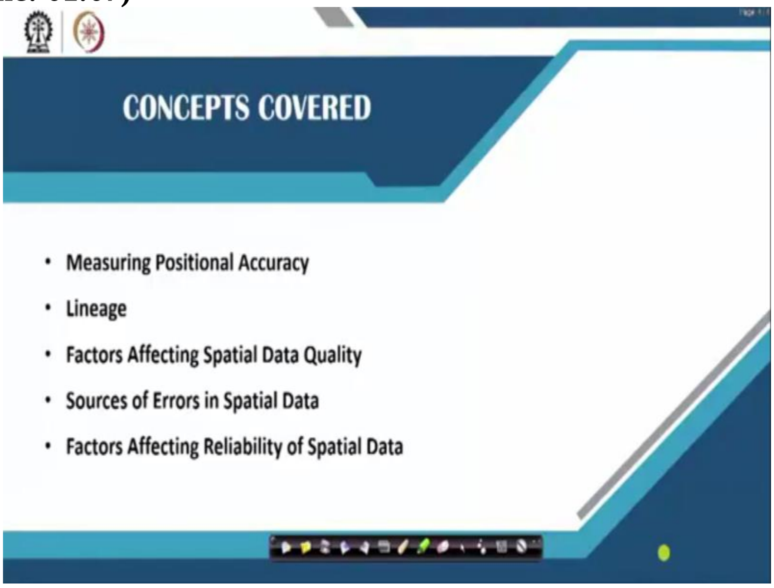
Geographic Information Systems
Prof. BHARATH H AITHAL
Ranbir and Chitra Gupta School of Infrastructure Design and Management
Indian Institute of Technology – Kharagpur

Lecture – 22
Positional Accuracy and Source of Errors

Hello, Namaste, I am back with the second lecture of the module 5, where we were discussing about the GIS data quality. I hope all of you have understood what do you mean by a data quality how data quality is important for your data, but now we will go into much more details of what do you mean we have positional accuracy. So, that is also very important, we discuss partially this in the previous class.

But we will also look at more aspects in this particular class and sources of error. So, we discuss some errors that may creep in, but we did not discuss what are the source of error. So, we are look at sources of errors, how do we rectify it also in the next 2 lectures.

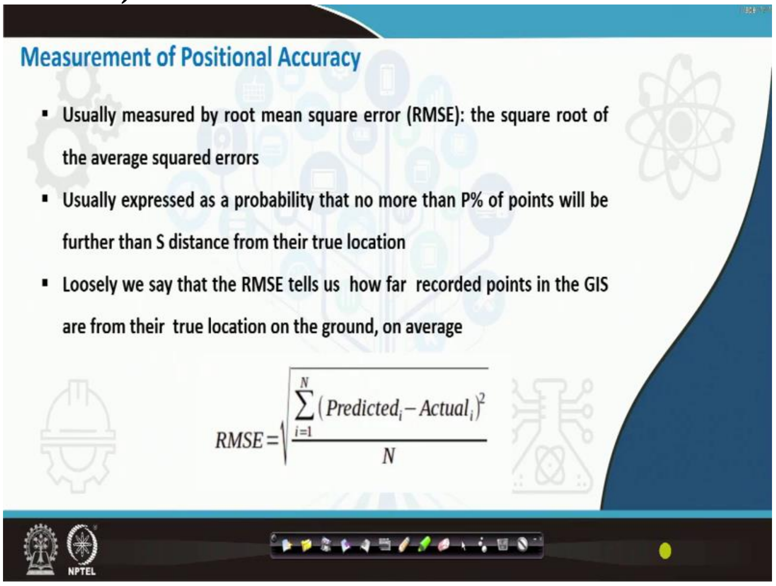
(Refer Slide Time: 01:07)



And, as I said, the concept that I would cover in today's class is measuring positional accuracy how you measure a positional accuracy, lineage, then factors affecting the spatial data quality sources of errors in spatial data and factors affecting the reliability of spatial data reliability is extremely like your completeness and logical consistency reliability is extremely important in terms of your spatial data.

How reliable is your data, so, you have to give an idea to the user that how what is the reliability aspect in your spatial data. So, that is extremely important when you are actually putting out this data in a public domain that can be used by many people. So, if that has to be done, then you are then the reliability aspects has to be considered in a proper way.

(Refer Slide Time: 01:53)



So, when you are looking at the measurement of positional accuracy, this is normally measured in times in terms of root means square error, the square root of the average square errors that is if you have predicted error and actual error. So, it is the root means square of the average errors that we that is there in the entire data set or the entire data layers that you consider that is how it is measured.

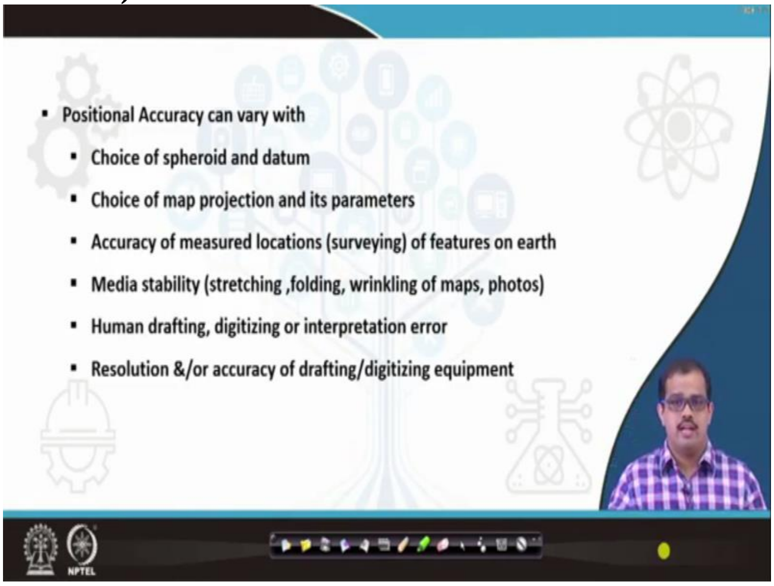
So, it is usually expressed as a probability that is not no more than P% of the points its farther than S distance from their true locations. So, P number of P% of points will be farther from the distance is S from their true location. So, how many is displaced from the true locations loosely, we say that RMSE tells us how far the recorded points is in the GIS are from the true locations on the ground.

For example, this had to be your location. When you have mapped it in the database, it has shifted to this location. So this is the shift that you will see in your true locations. That is what measurement by the RMSE. So once you calculate the RMSE from the real world, to your data model, then whatever data that you have recorded can be easily shifted by so many points or so many units.

So, that is how you improve your data, how you improve your positional accuracy. For example, when I when I am actually looking at a satellite data, the first thing is we go to the field, collect the data, we know what is a position of that particular data come back, tie the position to the particular information. So, for example, intersection of a maybe road a building that is already there from a very long time, very easily identifiable building.

So, all of these are the information that you have to collect. So, once you know this, tie the information to the real world coordinates or the real world information that you have already I have got. So now you can compare what is the error so with that, shift the particular point so that you almost have 0 pixel error or 0 error in terms or 0 RMSE. If that you are can archive then you have the best positional accuracy for a particular data.

(Refer Slide Time: 04:25)



So, it can also be there are various things that you have to consider when you are looking at personal accuracy, the first thing first and foremost thing is put it in the same coordinate system look at the spheroid and data it has to be same if it is same, then there is no issue in terms of I mean most of the maps are overlay. The first thing that you have to look at once both of them are matching them overlay.

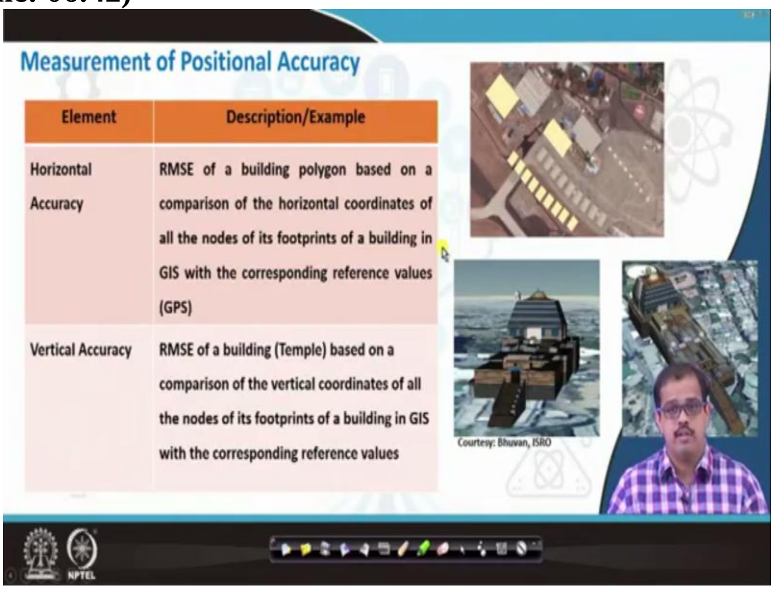
The next thing that you have to look at is what is the kind of measurement locations accuracy that you have considered, if you have a considered exact measurement locations accuracy, then you know where that particular point has to be. So, you can shift a particular point or if it is extremely accurate, then you can maintain the data set as it is then the media stability. What kind of stretching, folding, drinking of maps or photos we have a plan if you have a physical map.

Then human drafting or digitizing errors, this also can be easily found out. Nowadays, there are huge numbers of tools that they have developed in terms of understanding these interpreting errors. So, using those tools most of these errors can be either removed or rectified then you have resolution or accuracy of drafting or digitizing equipment. So, when you are actually looking at a drafting equipment, you know, what is the accuracy of that particular equipment.

So, that is extremely important in terms of looking at drafting or digitizing accuracy. So, if you have understood all of these in your positional accuracy then your data is quite accurate in terms of measurements. So, please keep in mind. The first thing that you have to look at us, what is your coordinate system, what is the projections that has been used, whether they match the data or that is a reference point that they have used?

And what is the accuracy of measurement on the real world your data model that you have to probably look at other things are digitizing errors that normally if you are now digitizing now automatically you can look at it without any issues. But these are the issues that you have to keep in mind when you are actually looking at errors the major errors.

(Refer Slide Time: 06:42)



So, let me give you some examples. For example here the first thing is a horizontal accuracy. So, when you look at this RMSE of a building polygon is based on comparison of horizontal coordinates. That is, if I am trying to see this polygons, it is dependent on the horizontal coordinates and all the nodes of the footprints of a building in GIS with the corresponding reference value on the ground that may be through your GPS or any other ways.

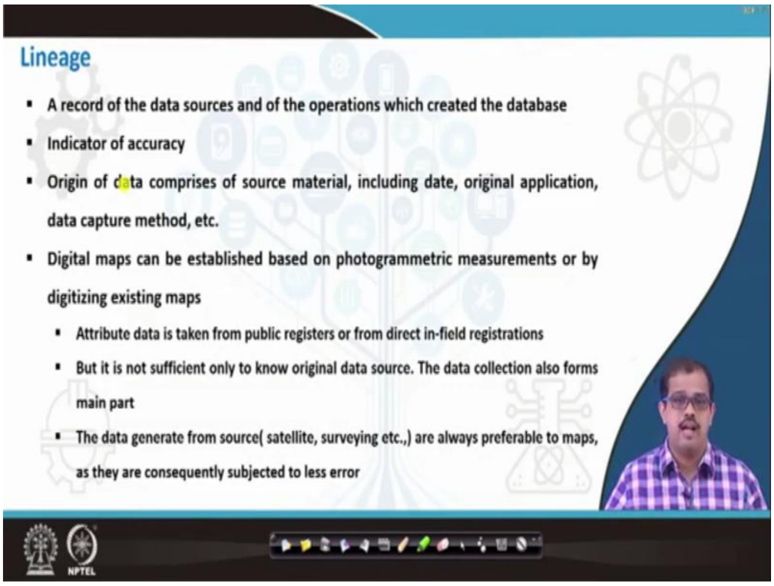
So you have to look at the corresponding points here. Look at the horizontal coordinates first. Of all these buildings fine or look at the sample number of these buildings then look at the notes of footprints are the coordinates of those notes are the footprints to the real world match it to the real world you know what is the difference, if there is certain difference, let us say a difference of 2 meters in a image that have time, so shift your image by 2 meters.

So, correct your image by 2 meters or your whatever the accurate measurements that you have already done, those measurements will be will be having an additional of 2 meters. So such event such thing has to be taken into consideration when you are looking at positional accuracy. Vertical accuracy vertical, most of them may not need a vertically accurate data because most of them work on a horizontal data.

But when you are looking at a vertical accuracy then it is extremely important for example when you are looking at this, you have a building here so this when you are looking at comparison of vertical coordinates of all of the notes of the footprints of the building GIS with the corresponding reference values. So, there are certain values which you can consider to be as a reference for this building.

So, you have to look at those reference values. Also, there are certain measuring instruments which can be used to measure the height of the building. So, if you have those measuring instruments, so, you can measure the height of a building and use it for such analysis. So, in order to see whether it matches the real world phenomena.

(Refer Slide Time: 09:05)



The next thing that is that we have to understand this lineage, how the record of that particular data has been created in the database, then how it has been transferred, how it has been used, applied and now used by someone else. So, how it has come over age, so, that we have to look at in terms of lineage. So, it is also an indicator of accuracy, how accurate is your data if your data is completely old without updation and you are trying to find out today's scenario of maybe growth.

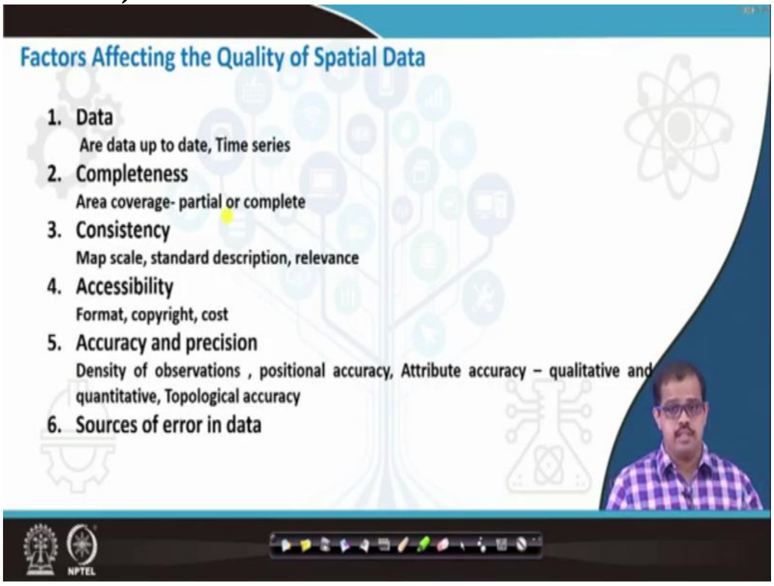
So, you will not be able to understand it, because your data would not give you exact point of in that particular map. Then origin of data comprised of source materials, including date, original application data capture method this have already explained in my previous lecture, digital maps can be established based on photogrammetric measurements or by digitizing the existing maps.

So, the best method to get accurate values is using the photogrammetric measurements. And if you are not very good with photogrammetric measurements, then look at digitizing existing maps. So attribute data is taken from the public registers often direct in from the field. So, this source has to be mentioned in your data. But this may not be sufficient only to know the data source, but also the data collection, how the data has been collected.

Even if that data has been collected from certain place. So, how that particular data has been collected to represent in that place has to be known. How the method of preparation of that particular data is also on the point that you have to look at the data generated from the source whether it is satellite surveying, etc., are always preferable to maps as they are consequently subjected to less errors.

So, you prepare it as maps and because of which you can look at what is the accuracy of that particular map and then use it in into your database.

(Refer Slide Time: 11:11)



So, there are certain factors affecting the quality of the spatial data also. So, when you are looking at this data itself as a first quality that is affecting that spatial data for example, are all your data up to date or not? For example, that is what I mentioned in my previous slide as an example, if your data is 10 -15 years old and you are trying to model today's phenomena, it will really not capture certain phenomena.

Let us say, we are capturing your phenomena how urban growth is happening maybe in a particular city. So, what you try to do is that first thing that you take here is you have let us say satellite data of 1990s and then you have a satellite data of 1995 and a satellite data of 2000. You have extracted the land use. Now, what you want to do is that you want to know what is the growth that is there in 2019.

But, without understanding what is a growth between 2001 and 2018, it may not be possible to find out what is a growth in 2019. Why, because, for example, a particular city then maybe an industrial region that has been set up in 90s there may be and IT hopes that have come up in 2000 and there will be certain agents that have fuelled the growth in 2000, 2005 etc, the growth may be higher spikes or lower spikes.

So, if you can understand this kind of dynamics, then only you will be able to modulate for 2000 otherwise, that particular area where the growth may be in a higher spike or maybe in a lower spike will not be captured effectively by your model and that leads to an issue in terms of understanding, then the completeness is your area coverage complete or partial? When I also say this your completeness in terms of also collection of data.

You should also complete you should also have an heterogeneous data features that has spread all over the data region it should not be only in particular region for example, if you are looking at urban area and the buffer your data collection should be both in the urban area and also in the buffer. If you are only collecting only in the urban area and not in the buffer, then your data is actually not complete incomplete.

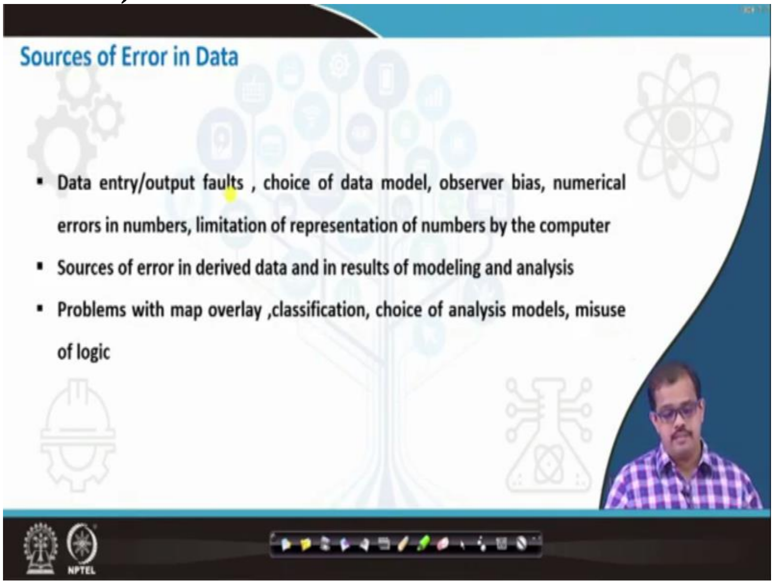
The way you are collecting data is actually inaccurate because it is not representing the true world phenomena. Consistency of your data is another important factors in spatial data quality. So, when I say map scale, the standard description the relevance the way of it the way it is represented so all of this should be consistent or when you digitize it, make it consistent.

So then it would be much easier for you to look at the data quality than accessibility for example, the formatting the copyright, the cost involved, so all of these are some factors that affect the accessibility, then accuracy and precision. I say said word data may not be both accurate and precise or it may be accurate and or precise, depending on the situation it is send.

So look at that accuracy and precision in terms of observations density of observations, positional accuracy, attribute accuracy, qualitative and quantitative both then topological accuracy. So, when I say positional, it is almost quantitative and when you are looking at accurate accuracy it as more of qualitative, sometimes its quantitative. So, look at both qualitative and quantitative aspects and topological accuracy.

So, if you have looked at all of these most of your data is will be quite accurate in terms of representations. Sources of errors in data if there are certain source of errors in data the way it is captured way it is measured, so, that may also affect add as a factor that is affecting the quality of a spatial data.

(Refer Slide Time: 15:06)



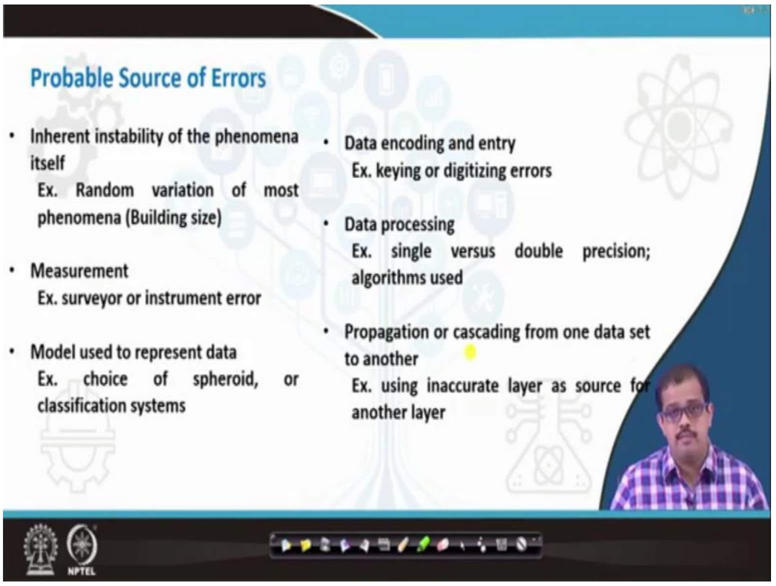
So, when we look at that source of error, the first source of error is a user that is through data entry and output faults, that is a first thing that a user makes, so the first source of error the major source of error that I would assign is the source of error in a data. So, also the choice of a data model this is extremely important in terms of when what kind of data model you are using.

So, if a user is using a wrong data model, then I sure to then your entire spatial data is not going to be propagated in a proper way then the bias. So, that bias should not be there, it has to be neutral agreement between any phenomena in order to assess that particular data properly numerical errors in numbers maybe it may creep in many times, because of a huge data sets, limitations of representation of numbers by a computer.

This may not be valid today, but in certain old systems, yes, it is there are certain issues limitations when you are looking at limitation of representation, then you have source of error in derived data and reserves of modelling and analysis. So, this is also extremely important in terms of understanding the error and problems with map overlay classification, choice of analysis models.

So, I have spoke about overlay classification already, then the misuse of logic, logical consistency if it does not have, then you have a source of errors and data.

(Refer Slide Time: 16:46)

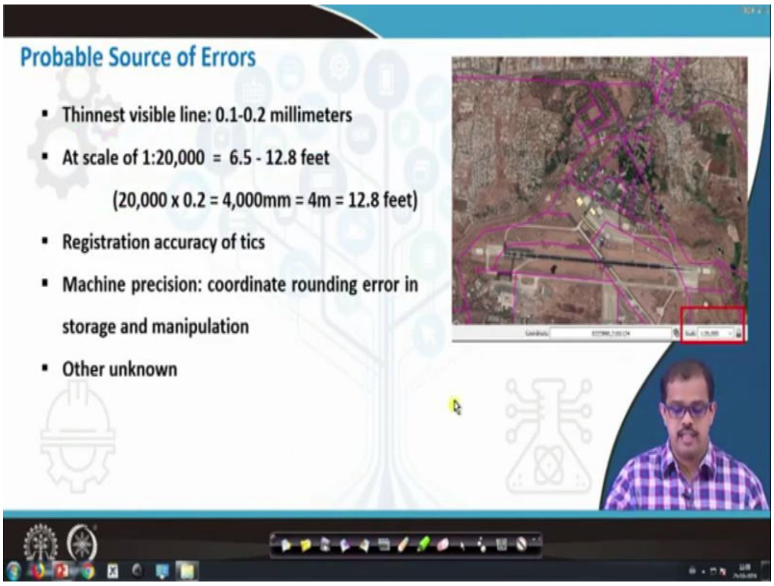


So, please look at all of these, then I have also given some probable source of errors, these are inherent instability of a phenomena itself for example, random variation of most phenomena, building size if you are looking at 5 years and if there is some issue as changing building size and you are taking that as a reference point. So, probably that also can creep in as an error.

Measurement instrument comes as an instrument error or a surveyor error. It will be model used to represent data that I have already spoken whether what kind of spheroid it has been considered what kind of classification system it has considered then data encoding data processing and finally, the thing that you have to look at in terms of data error is propagation or cascading from one data set to other.

So, if you have an error other in data one data set, it actually propagates to all of the data sets because of which the continuous in the data set may increase, and that may lead to more errors.

(Refer Slide Time: 17:47)



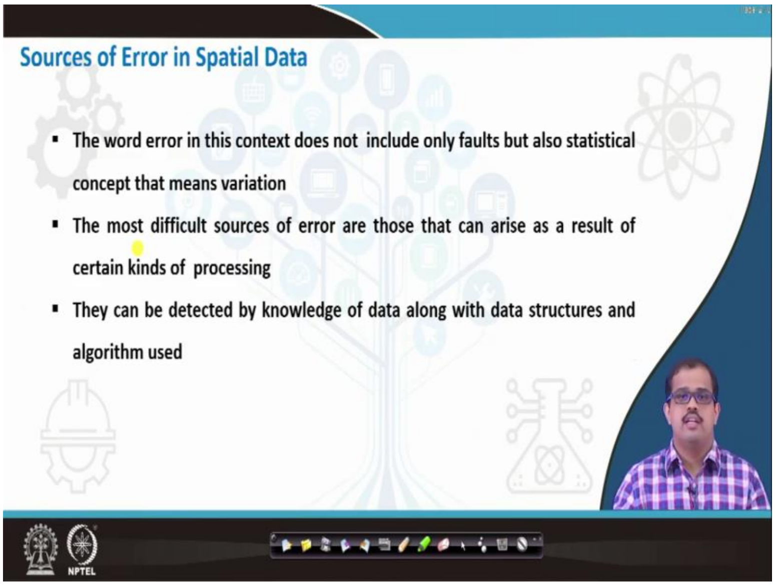
So, to just give you an example here, if we consider these lines that have digitized on this particular map here, so I will just use some other color. I have digitized all these maps here. So, if you look at this, this is and 1:2 it is 1:20,000 map. So if you are looking at this particular map, if this is the error, and the finish line that you can see is 0.1 to 0.2 2 millimetres.

So, at a scale of 1 20,000 if we look at it is almost close to 12 feet or 12.8 feet to be exact. So, which means to say that you can look at 4 meters that is equal to 12.8 feet yet a small deviation in this will give you 4 meter error that 4 meter error on the ground and 12.8 feet or sorry 12 4 meter error on the map and 12.88 feet are on the ground. So, which means to say that the registration accuracy would be going wrong.

So, when you are trying to measure the distance when you are trying to measure the building area etc., then you are actually making a mistake. So, when you are looking at machine precision operation, the coordinate rounding error is in the form of storage and manipulations many a times and some other unknown arrows can also occur. This has some errors that have put forward but when you are actually working.

You will understand there are a lot of other challenging errors that you may see when you are looking at the spatial data.

(Refer Slide Time: 19:27)

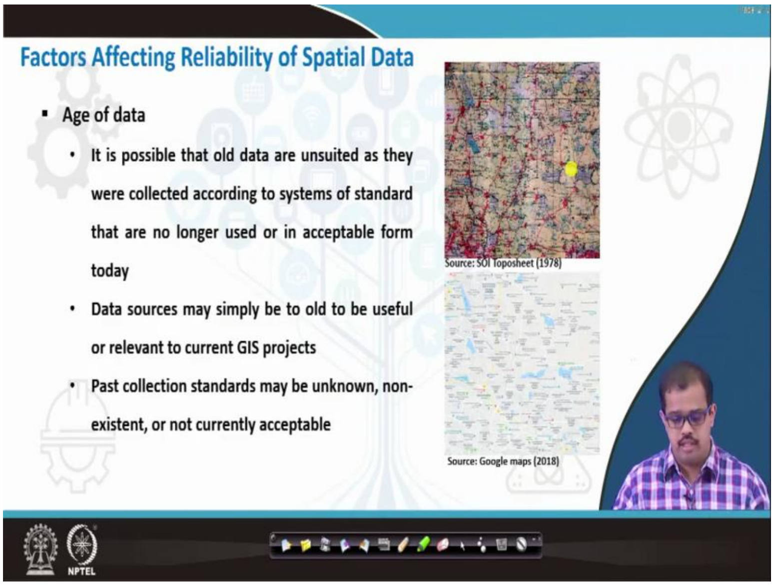


So, when we are looking again at this thing, the word error in the in the in this context does not include only faults, but also statistical concepts. For example, the mean variations, so, that is also an error in terms of the spatial database you have considered the entire data if consider mean of that data. So, what is the amount of error that it can creep in and also is under the kind of error.

The most difficult source of error are those that can rise as a result of certain kinds of processing, many times we use certain processing when we do not have an access to high computing capability systems. So, we use an approximate model of our processing methods in which our computational not intensive that can also provide certain errors into the entire data model.

So, they can be detected by the knowledge of data along with the data structures and algorithms that are used. So, if you understand all of these the algorithms and data structures and the data that have been used, then these errors can be easily removed from your database.

(Refer Slide Time: 20:44)



So, the factors in case you are looking at reliability of data which is extremely important in terms of spatial at how reliable is the data that has been generated by you are in case you are actually transmitting that data for different users to use in different applications, then the reliability is extremely important. So, how do you find out that it is the first thing that any time that you look at need age of data.

If it is possible that old data are unsuited as they were collected according to the systems of the standards that are no longer used or an acceptable form it can be this issue or it can be the old data that may not be giving the information that is there today. For example, in 90s, we had the data at a resolution satellite data resolution of 72 meters and the 36 meters. But today, we have a data at 1 meter resolution at a centimetre level resolution.

So, if you are comparing a centimetre data with data which is 72 meters or even 1 kilometre data, then you that is where you are the first reliability goes as far as your data measurement is concerned, whatever you are measuring measure at so, that the sources are actually same. Basically, without the old data which may not be effective enough in capturing the phenomena.

Then data sources may simply be old to be useful or relevant to current GIS product projects as I explained in the past collection standards, so, this is extremely important in today's context, the standards of either collection, representation depiction, all of these have been standardized to a very large extent. So, past collection standards may be quite unknown in terms of or maybe erroneous.

So, in order to have our understand have a better data model, it is better to actually avoid old data then and create a fresh data and which may actually affect your entire data model. If these factors are considered your data may be quite reliable in terms of accuracy of that particular data. For example, here have taken up a survey of India toposheet. And here is the Google.

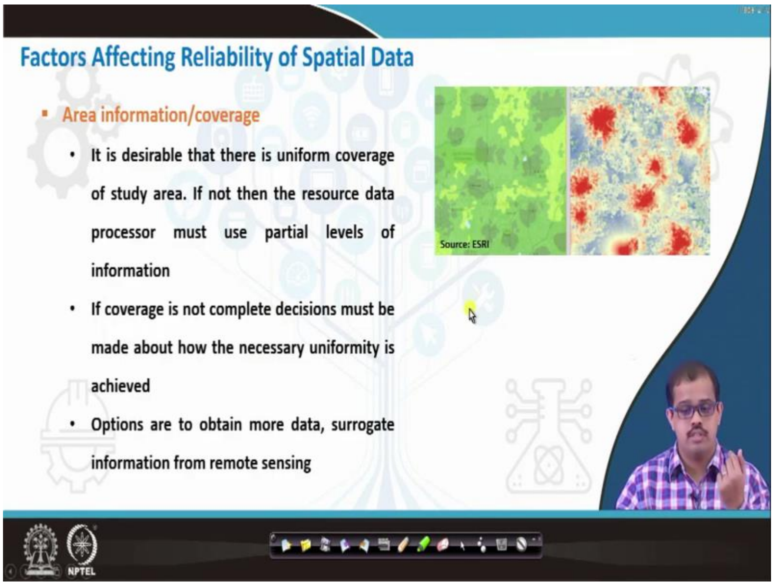
Now Google maps can go up to a distance of you and get information at a 10 meter distance and the accuracy of that information as per many sources that has been represented in research that is plus 3 to 2 plus 6 meters. Vary upon across regions to at least 10 meters. So, now if you have that accuracy the horizontal accuracy is about 10 meters. And when you look at the SOI toposheet that is there that has no that is normally not updated very frequently and has been in time from 1970s onwards.

So, if you are trying to look at these to then these is old information may not be logically important in terms of analyzing this particular data that is there in today's context. But this can provide a vital source of information for your data model. If you are trying to find out what was there in 1970s and what is not there today. So, in terms of extraction of what was there, but then corrected according to today's standards.

Maybe the best model that you if you are looking at the old data versus the new data, the corrected data can be then used to compare with the new data. So, correction levels of what kind of correction can be done can be very well known for a GIS user. Anyone who has some experience of handling GIS data can easily start looking at such phenomena where they can easily extract information and then convert it in such a way that it becomes a more reliable data in terms of accuracy.

So, you have to look at 2, 3 different sources of the same data. Once you have understood the source, then you will be able to understand how reliable this particular data.

(Refer Slide Time: 25:05)

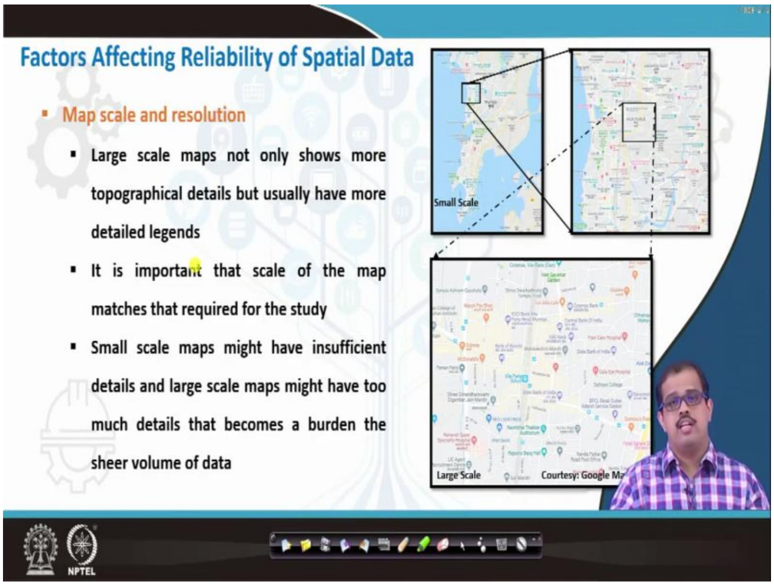


Then the area information or coverage, so it may be that most of the data that you have considered as uniform, for example, if you look at some of my application, some of the areas that I have considered for example, if I have working on Kolkata entire Kolkata metropolitan region. So, if I have concert 10 kilometre buffer, so many of the places many of the years I could not at least 10% of that data of the spatial data that is available in that 10 kilometre buffer.

So, what we did we try to always use a similar buffer neglecting that part of the region. So, if you have removed that entire part of the region for your entire course of the day, so, that would give you some meaningful information in terms of reliability. Otherwise, if you have in one year you have that set data set another one you do not have a data set you cannot compare it pixel by pixel understand what is the change.

So, always having the area information or change should be on the same polygon or the same region that you have considered. Options can be used to obtain more data surrogate information from remote sensing also.

(Refer Slide Time: 26:24)



So map scale and resolution so, this is also extremely important in terms of when you are you are looking at the reliability. So, whatever the maps are used may if you can try and get the large scale map it would be more important in terms of extraction of topological details, but, usually we have more detailed legends in these large scale map. So, look at large scale map. So, again let me repeat we understood in our previous session we have to me we have large scale map and if you have a small scale map.

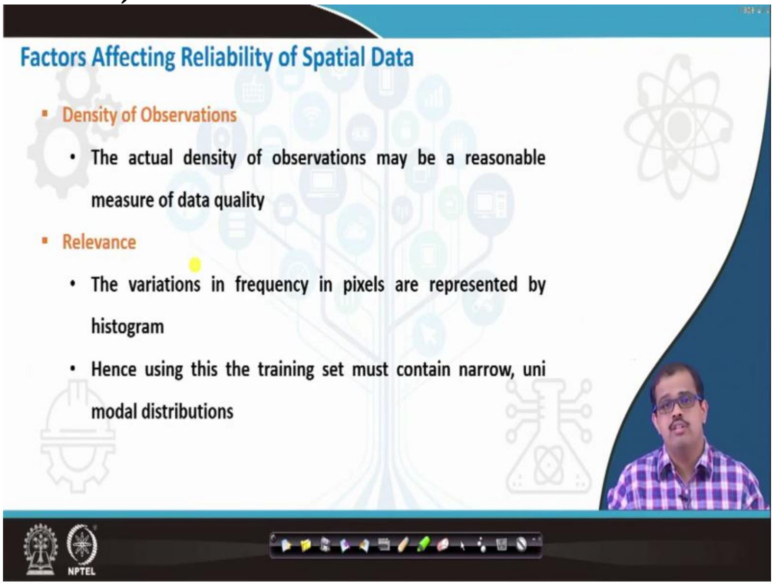
Large scale map gives you a large amount of details whereas, a small scale map gives you a very less details. When I say small scale maps these are 1:1,000,000 1:20000 maps and I say large scale map it can be compared to 1:1,000,000 1:10000 1:25,000 is a large scale map. So, keep this in mind. So, look at the information that is actually necessary, it is important that the scale of a map matches the required area of the study.

So, for example, if you are trying to study the entire subcontinent of India I mean entire India as a steady region and if you are trying to use 1:5,000 or 1:10,000 map may be foolishness, because when you are trying to process this entire data set. It may not work out in terms of giving you an accurate data or generating an accurate data. Instead, if we used data at if you use certain ways of generating this information.

Wherein you are using maybe 1:50000 maps or one even 1:20000 maps and it can be extremely eventful in terms of mapping the entire India and extracting its details. So, and small scale maps most probably as I said I have insufficient details. So, now, the user has to decide small scale map has insufficient details or it has very less details whereas a large scale map has shear volume of details.

So, the computing capability here was does not know computing capability is necessary very less computing capability that is necessary. So, all of these factors has to be considered before you look at what is the reliability of your spatial data.

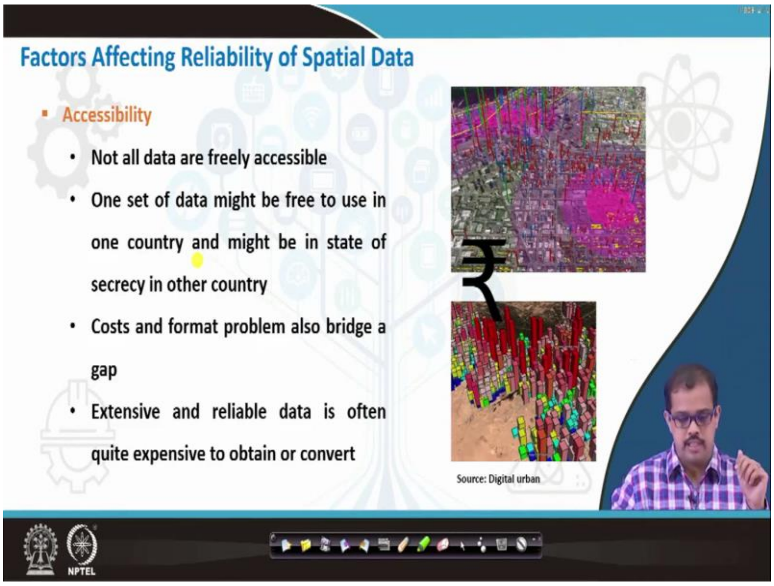
(Refer Slide Time: 28:49)



Then density of observations what is the density of observations that you have maybe a reasonable measure of data quality not always it is an important part, but they are there should be certain ways of analyzing the density. Then relevance the variations in the frequency in pixels are represented by histogram for example. So, this can be helpful in terms of looking at what are the training sets has to be considered what are different training groups that have to be considered, and what are the distributions.

For example, the model distribution that it can have in when you are trying to classify a particular data.

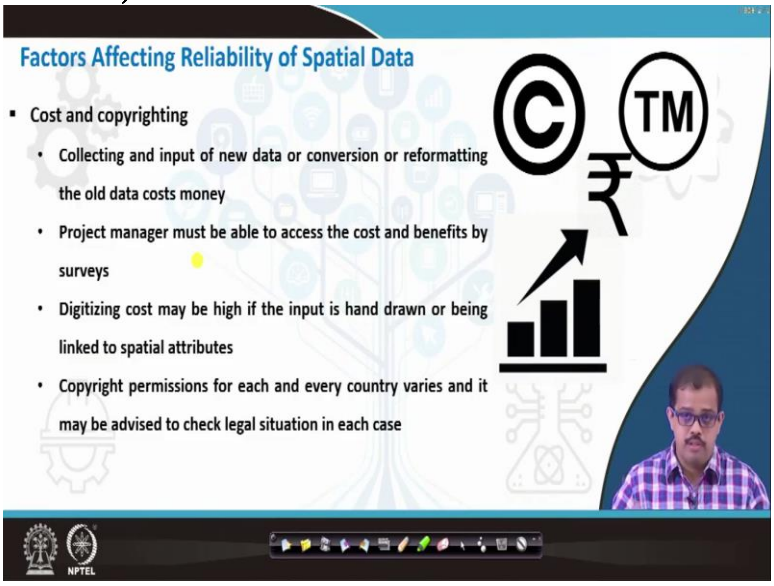
(Refer Slide Time: 29:31)



So, these are certain things that you have to look at. And the last thing that is very important as accessibility. So, as I previously said, not all data is freely accessible, one set of data might be free to use in the in one country and might be the state of secrecy in other countries. So look at what data at a specific country this user it maybe even between the state even between the corporations also.

Cost and format problems also bridge a gap. Now, the most data that India is producing is so most of them are the interoperability standards. So looking at this, this may not be a much of a big issue. Extensive and reliable data is often quite expensive to obtain or convert. Yes, this is the very important part. Most of the private generators who actually generate extensive reliable data are at extremely expensive in terms of obtaining this particular data.

(Refer Slide Time: 30:30)



And then you have the cost and copywriting so cost I have already spoke, the major thing is copyright permissions for each and every country varies and it may be advised to check legal situation in each of this country and most importantly the copyright permissions of the data that may be available publicly. But in order to use it in your own data model, it may need certain permission certain cost that may be involved in order to obtain those permissions is also important.

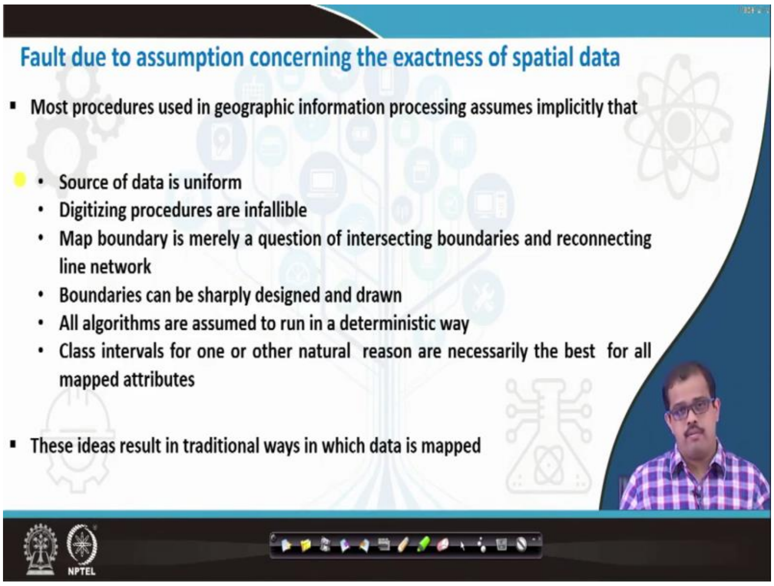
So without that, it may not be important to you or it may not be using to your database.

(Refer Slide Time: 31:06)

So, another error that may come in are further checking the reliability of data is a numerical error. So when you are looking at the computer processing capability or the ability to store and process the data at required precision that is also so, very important. So, if you are precision length is 6 points, so, if your computer is able to store only at 2 points, then that is where you are numerical error starts.

Then there is insufficient pressure and maybe lead to calculation errors also for example, if you are calculating the length of fraud our area of polygon, so, it may also creating certain amount of precision errors in the data, so, number of polygons will actually populate it to much bigger errors. So without that, it may not be important to you or it may not be using to your database.

(Refer Slide Time: 31:50)



So, then the last thing that I would speak in this particular class is the fall due to assumption considering the exactness of spatial data. So, we think that we assume basically we do not think but we assume that most of the sources of data that has been conserved as uniform and digitizing procedures that one may have use infallible. So, which has to be understood. You have to look at the data probably, once you look at the data.

You will understand how consistent or inconsistent that particular digitizing has been. So, when you look at most of the people who extract map boundaries, they look at only intersecting boundaries and reconnecting the line network, which may not be true in the real scenario, boundaries can be sharply redesigned and drawn. So, it has to be properly redesigned and drawn.

Whenever you are considering the boundaries, all algorithms are assumed to be run a deterministic way. All algorithms are considered to be true, all algorithms are considered to be the best so but it may not be really true in terms of real world scenario. So, that these are certain things that we have to use, when we are looking at how we are mapping in any of the traditional ways.

That are there, and how we are mapping in the digital race. So if these ideas are considered, then it would be the same traditional way of mapping and error that accumulates into the map when you are creating the map or a data output.

(Refer Slide Time: 33:30)

So, to summarize today's class, we looked at measuring positional accuracy, how different methods of measuring the positional accuracy, we looked at the lineage and how the data has been transferred over period of time, then factors affecting the spatial data quality errors in the data, or a spatial data then factors affecting the reliability. So, we looked at all the factors starting from how the generation happens and what is accessibility and the cost.

That is necessary. And then, probably in the next class, we will look at the pixel accuracy for example, when you are using a raster data what is accuracy measures? How do you calculate an overall accuracy? How do you calculate a kappa statistics? How do you say that whatever the land use that you have generated using a raster data, not a raster data model, you can satisfy the criterion of providing the accurate data. So we look at all of these in probably in the next class. So till then, have a nice time. Thank you very much.