International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.7 (2015), pp.63-78
http://dx.doi.org/10.14257/ijsip.2015.8.7.07

Handling Endogeneity Challenge in Big Astronomical Data

Sumedha Arora and PankajDeep Kaur

Guru Nanak Dev University (Rc-Jalandhar)
sumedhaaroa7@gmail.com, pankajdeepkaur@gmail.com

Abstract

Using Big Data in statistically valid ways is posing a great challenge. The main misconception that lies in using Big Data is the belief that volume of data can compensate for any other deficiency in data. There is a need to use some standards and transparency when using Big Data in survey research. Certain surveys that are based on the Big Data tend to generate more complications and complexities in data such as some important variables tend to correlate with some errournious data. This correlation of data with residual noise causes the endogeneity problem. It is to be solved as a fact the main aim of research work is answering question which could only be done when data is fully analyzed. Through this we can utilize all available information. This paper throws light on addressing endogeneity particularly to the astronomical data set and also provides solutions and techniques for handling endogeneity in the respective data set. Finally it couples big data i.e. whole data of sky with the time domain.

***Keywords**: Big data; high data degree; noise collection; data storage; incidental endogeneity*

1. Introduction

Astronomy is famous for acquiring, interpreting and systematizing large quantities of data. Like many other fields Astronomical data is gaining importance day by day. It is framed by rich data sets. These rich data sets are framed from advancement in telescope, detector, and computer technology .Content of continuous sky surveys is measured in terabytes and petabyte. Drastic survey of sky, stars, galaxies make use of large wavelength hence becoming the primary source for massive amount of data. Other sources that give rise to big data are numerical simulations. These technical problem ranges from many challenges including incidental endogeneity, heterogeneity leading to the need of developing solutions to deal with these challenges. If we look through the history of data, we see that single objects were used in individual studies and now in today's scenario we have to map the whole sky/universe systematically. Exploration of these real scientific discoveries are making data analyses methods inadequate to handle [1]. There are mainly two reasons for astronomical data experiencing great expose to data first being the development of new telescopes in India like (LSST) which can image enormous region of sky and Secondly sensitivity of the detectors follows the Moore's law that is large images increases with pixels. So the main concern here is how to search and how to synthesize that output.

The two main issues driving current data challenges in astronomy are (i) growth rate of data i.e. over the past 30 years we are able to build the telescopes that are 30 times larger than the previous telescopes moreover their detectors are 3000 times more powerful in terms of pixel rate. This ultimately results in exponential increase in the data, which means data rate is being doubled every year thus becoming difficult to capture and analyze. Increase in heavy range of big data effects the entire data sets of the universe. Various data sets ranges from educational data sets to astronomical data sets. This further gives rise to the certain challenges. One of the major challenges of big data include incidental endogeneity. These challenges differ from each other and require new paradigm to solve them. We have reached the phase where storage of big data is not as big an issue as to analyze it, the main issue is to build a cloud that can host large data sets of astronomy in order to handle these massive data sets. So there are certain technical problems which range from database design and federation to data mining and advanced visualization that leads to the development of a new toolkit for astronomical research.

So starting with the related work of Keller and Masts coming to the endogeneity problem in section III, paper discusses the astronomical data, its characteristics, its relation with big data and try to address and solve the endogeneity problem in it. It also gives the appropriate solutions to solve the endogeneity problem. Finally it gives the data measurements in VIIth section and also the results to verify the experiments discussed in the paper.

2. Related Work

2.1. Kepler

Kepler measures the light of 170,000 stars very precisely at regular intervals looking for these dips in light that can indicate presence of a planet. The area that was sampled was not very large, though it was a small patch of sky, the area they were sampling was done after every 30 minutes which created lots of data. It was opening up the time domain as here you get new images after every 30 minutes, which creates chaos. So Kepler was the solution to these kind of surveys. Through Kepler the ability to have a close look of the object in the sky was developed. But if we want to have a deeper look of sky then we have to move on to the Large Synoptic Telescopes.

ALMA, the Atacama Large Millimetre Array, very soon it is going to release its first data release and its raw data ranges forty terabytes a day it can produce more data than we now have on the entire internet, in other words it can flood our internet with big data, which can further give rise to challenges.

2.2. Mast

The Multimission Archive at the Space Telescope Science Institute (MAST) archives [3] a variety of astronomical data. They find the way to analyze and access big data with the help of some correlation tools. These tools allow the users to search all the archived data and to efficiently use it. Even the preview of images can be obtained through this method. Based on the archival nature of the requested data, MAST can provide access to the requested data in different ways which includes intermediate disk staging and direct web based downloads. MAST [4] data rate can exceed ten Terabytes, including all the links to archival [5] data.

3. Major Problem in Big Data-incidental Endogeneity

In recent models with high dimensionality is earning greater importance in many fields like science, engineering and humanities. The number of regression keeps on increasing in these kinds of fields. Incidental endogeneity refers to the relationship between the variables and the sample terms. It occurs due to the results of error measurements and due to the omitted variables. Some predictors can correlate with the residual noise. It arises mainly in big data due to the following reasons [18]:

I. Scientist these days are inspired to collect as many features as possible so this increases the possibilities that few data get correlated with the residual errors and external noise incidentally.

II. As the big data is collected from the different sources, this increases the possibility of selection bias and errors. Hence it give rise to potential incidental endogeneity.

Moreover a variable is called *endogenous when is* correlated with the error terms else it is named as *exogenous.*

3.1. Method to Handle Endogeneity

It is equally important to develop methods that can handle endogeneity.

- **Penalised Least Square Method**

This method was not able to achieve the variable selection consistency. It is due to the reason that it does not contain least square loss function this method will fail even if the endogeneity is present in unimportant regressors. Though it can select many unimportant regressors.

- **Novel Penalized Focused Generalized Method**

FGMM provides an extra filter that excludes all the endogenous predictors [19]. Through this method we can not only select the important regressors but we can also eliminate many unimportant one's. 'F' in FGMM has three meanings (i) Focus on correct moment equations (ii) filtering of endogenous predictors, and focus on selective model. consistency. FGMM consists of the oracle property even in the presence of incidental endogenous predictors that provides the solution that is globally minimum [20].

- **Insertion of Indicator Function**

The massive dimensionality problem can be handled if we include the indicator function to it. It not only solves the endogeneity problem but also it helps in reducing the high dimension of the data .Despite of its high advantages it still contains few disadvantages *i.e.*, it makes FGMM functions unsmooth.

4. Astronomical Data and its Characteristics

Astronomical data is a part of scientific field which concentrates on the study of astronomical objects. It also deals with the phenomena that often changes with the time. For example pulsating stars, asteroids, eclipses, galaxies. Nowadays, there is a 24 hours survey of sky with the help of big advance telescopes. Because of this the data volume is increasing drastically. In future there is a scope of capturing even a digital movie of whole sky. More perfectly we can realise the work of Large synoptic Survey Telescope, it can generate data streams at rates of 2 Tera bytes per hour, it can also generate 150 petabyte imaging data sets of whole sky. This give rise to the big data and subsequently with big data comes the incidental endogeneity problem. so there is a need to remove this challenge in this particular data set in order to maintain the technical development of INDIA. So in the following sections of the paper sources of endogeneity in astronomical data are discussed and a relevant solution to handle these challenges is provided. It also provides different paradigms for astronomy like that of statistics, data mining, knowledge discovery, machine learning and computational intelligence.

4.1. Heterogeneous Data

Astronomical data are heterogeneous by nature in terms of both contents and the format. Astronomers these days are busy in the exploration of each and every region of universe that includes electromagnetic spectrum. Imaging data that is the most important part of the astronomical data. It captures the two dimensional picture of whole Universe within a narrow wavelength region at a particular instant of time. For processing of the imaging data catalogs are maintained. Each detected source contains massive amount of measured parameters which include different flux quantities, areal extant and coordinates. Coordinates are used to specify the location of astronomical sources in the sky [1].

4.2. Complicated Data

The Multidimensional Archive at the Space Telescope Science Institute (MAST) archives [3] a variety of astronomical data, they find the way to analyze and access the big data with the help of some correlation tools. These tools allow the users to search all the archived data and allow them to use it efficiently. Even the preview of images can be obtained though this method. Based on the archival nature of the requested data, MAST can provide access to the requested data in different ways which includes intermediate disk staging and direct web based downloads. MAST [4] data rate can exceed ten Terabytes, including all the links to archival [5] data.

4.3. Imaging Data

Imaging data is the one of the constituent of astronomical data. Many two dimensional images are captured during the survey of sky through large telescopes[2]. There can be the images of whole sky. The images can be of any dimension including 2D and 3D images.

4.4. Relation of Big Data with Astronomical Data

Astronomy is experiencing continuous exploration of data mainly due to the two reasons that are

There is a great development in building of the telescopes that can image maximum portion of the data.

- Images captured are denser with the pixels. So now the main job of the astronauts is that they have to access the quality of the data from the raw data. They have to filter the quality of data from the vast amount of raw data that they have captured.
- So they have to do lots of analysis before they begin with the actual research work. So the real struggle is how to search and synthesis with this.

5. Challenges to Astronomical Data Set With Respect To Big Data That Causes Endogeneity

5.1. Endogeneity in Astronomical Data

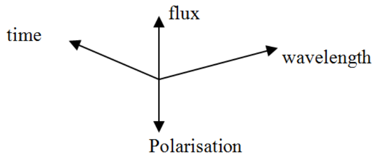
Petabyte of the data that is captured every day is not the major issue, the major issue would be the wastage of certain bandwidth and time. For example Kepler as discussed in the previous section measures the light of 170,000 stars very precisely at regular intervals looking for these dips in light that indicate a planet is present. Though the area of sampling is not very large, the major concern is that the survey is taken after every thirty minutes. Large Synoptic Survey Telescope (LSST) are capturing the movie of the whole sky that we can point a particular place in a sky and on the very another day we realise that there was nothing there. This is the main reason of endogeneity in the astronomical data sets

5.2. Changing Astronomical Style of Observational Astronomy

Old Ways Present ways Future	Now	Future

Pointed, large homogeneous observations (~ MB - GB)	Large, homogeneous sky surveys (multi-TB, ~ 106 - 109 sources)	Multiple, federated sky surveys and archives (~ PB)
Small samples of objects (~ 101 - 103)	Archives of pointed observations(~TB)	**Virtual Observatory**

As it is clear from the above table that with the advancement in the technology we are getting the complete physical picture of the Universe. This results in increasing the complexity of data, which further increases the demands for data analysis, visualization and understanding ,so there is a need to create a dynamical, web-based research and interactive environment for the new astronomy with massive data sets, this process is called VO [8].



As we can see that measurements are taken along each axis. And each axis are characterized by its position, sampling, extend and resolution. So the data capturing rate keeps on increasing, giving rise to incidental endogeneity.

i.e. during data capturing from such a big raw text few variables get omitted and few get mixed with errors ,so it becomes very important to remove this problem, the following section of the paper deals with the endogeneity problem by performing certain experiments on data sets like Galex and star.

6. General Techniques for Solving Open Problems for Large Scale Astro-statics

6.1. Developing Effective Statistical Tools and Algorithm for Dealing with Big Data

Algorithms that are commonly used for operations in astronomy are Euclidean minimum spanning tree, neighbours, n-point correlation, kernel regression, kernel density estimation(KDE), comparing of the spatial structure of two data sets, are done by N-point correlations *e.g.*, luminous red galaxies in the Sloan digital sky survey [53]. KDE is used for comparing the distributions of different kinds of objects. Map reduce model can be used [60] for massively parallel implementation of K mean.

6.2. Implementing High Performance Computing

We can make use of GPU (Graphical processing units) instead of CPU. It is relatively new method for parallel applications in which high calculations are needed. Other advantages of GPU’s are their low costs and good processing power. It is inherently parallel i.e. it is a new paradigm for highly parallel applications in which complex calculations are offloaded to the GPU.

6.3. Use of Astronomical Pipelines

We can make use of astronomical pipelines by using cloud computing, through this computing resources can be scaled easily in accordance with changing work load. This can make the processing fast and quick. Many instructions can be executed simultaneously without waiting for one to complete.

7. Astronomical Data Mining

Astroinformatics is an emerging discipline which is crucial to deal with the data sets produced by new generation of instruments, sensors and computer simulations. Astroinformatics is triggering a true methodological shift which can be better understood by taking in account data quantity and data complexity. The parameter space can be seen as the N dimensionality manifold defined by the astronomical observables. Each observations can be expressed as a string of numerical measures – *e.g.* Right ascension, declination, epoch, flux in a given band, morphological type, etc. – which projects in N as a point or as a hyperplane or hypervolume of lower dimensionality. Day by day the new dimensions are added in the space due to the invention of the new kind of instruments. For example, quasars were disentangled from stars when the radio flux dimension was added.

7.1 Dame

DAME—Data Mining and Exploration
There is a need to perform the following things

- to provide the users with an easy access to both methods and computing power;
- to identify and implement better and faster algorithms;
- to minimize or reduce the data transfer by moving the programs rather than the data.

There by the DAME (Data Mining and Exploration) program is based on a platform which allows the scientific community to perform data mining and exploration experiments on massive data sets. It is done using a simple web browser. DAME offers several tools which can be seen as working environments such as clustering, classification, regression, feature extraction, etc. Furthermore the DAME infrastructure offers the possibility to extend the original library of available tools. This is done by allowing the end users to plug-in and executes their own codes in a more simple way. It also allows to upload the programs without any restriction about the programming language, and helps in automatically installing them through a simple interactive procedure. Moreover, the DAME platform offers a variety of computing facilities, which are organized as a cloud of versatile architectures, from the single multi-core processor to a grid farm, automatically assigned at runtime to the user task, depending on the specific problem, as well as on the computing and storage requirements.

7.2. Drawbacks for Data Mining

First of all the Data Mining praxis requires, a lengthy fine tuning procedure which implies ten’s and sometimes hundreds of experiments to be performed in order to identify the optimal method, the optimal architecture or combination of parameters. In the case of MDS, even the decimated data would still pose serious computational challenges which could not be solved by most users. Second, astronomical data mining is extremely heterogeneous. It is unthinkable to move such data volumes across the network from the distributed data repositories to a myriad of different users.

8. Experiment

8.1. Galex

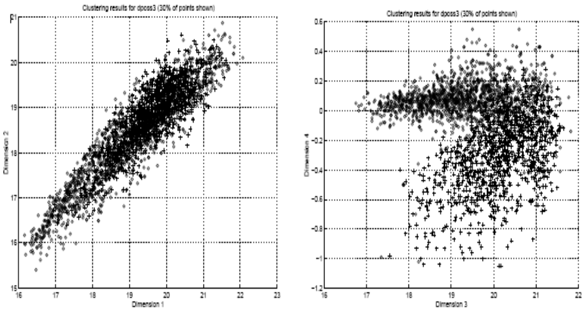


Figure 1. Two dominant clusters are shown, circles are representing (stars) and crosses (galaxies). The graph on the left shows a parameter projection in which the two classes are completely mixed. The graph the right shows two classes separate classes.

Galex is one of the major deal because it includes whole sky ultraviolet missions. Here the stress will be on the whole sky because measuring the whole sky ultraviolet sources is more data intensive rather than stressing on or zooming the single source of data. This is the major source of data archives at the time. The range of data it produces is gigantic. It shows all the object that had particular colour it also describes all the sources from the certain position in the sky.

The main objective is the discovery of clusters of stars or galaxies in the space. It is done by the proper utilization of full information that is available in space. The another objective can be automotive approach to galaxies classification [9], different clusters are dissimilar. Survey results differ every day, major cause of big data for example as depicted in fig 1. one day survey will depict a group of combinations of stars and galaxies, other day it will show that the stars and the galaxies are grouped differently .Encoded as circle are star and crosses are galaxies. Clusters and classifications help us in finding the new objects in a universe.

8.2. Stars

Another data set can be Stars:

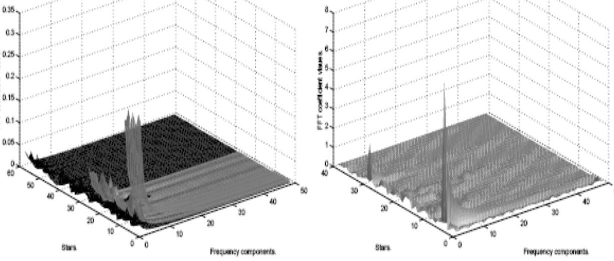


Figure 2. Left graph each cluster is represented as a smooth surface, i.e. stars are similar in clusters, right fig is showing distinguishable structure depicting that stars in different cluster are dissimilar.

If we capture the image of whole sky during night again we get the different data every time, for example graph on left hand side shows the plot of FFT coefficients of stars in 3 clusters and in right hand side image is of the stars that are captured randomly in the group, again the data varies with time which ultimately effects the data rate activities. These types of surveys create the time varying issues of data. As seen in the Figure 3, If we notice the left graph each cluster is represented as a smooth surface, depicting that the stars are similar in clusters where as graph on right hand is showing distinguishable structure depicting that stars in different clusters are dissimilar.

Techniques for dealing with the time varying data:

- **Clustering**

Clustering is use to detect rare and unusual object, that can be used for further investigations. It is done in order to refine the astronomical classification of objects in more objective ways.

Clustering techniques [8] are used in the case of mixture of data to find the group of interest, to find the descriptive summaries. It is also use to certain estimation for the large set of groups. Our main aim of the clustering technique is to find how may distinct types of object are found in the massive data. For this implementation the best way could be to put the similar data into one group. As some people would want only stars whereas others may want only galaxies etc.

- **Classifications**

In order to maintain the catalog we need to classify the astronomical sources into different objects. Different objects are categorised into different classes like that of stars or galaxies. We can make use of various classification techniques that can be used for this task. The original Galaxy Zoo is a project for a galaxy classification, it helps to classify galaxies by shape. The different type of galaxies can be round galaxies, elliptical galaxies, spiral galaxies, irregular galaxies. The shapes of galaxies reveals there history of the formation.

- **Catalogs**

In order to process image data, catalogs are generated. As we know that each source can contain number of different properties and they may also include large number of different parameters like that of flux quantities. All the information can be displayed in space provided in a catalog.

9. Addressing Endogeneity in Big Astronomical Data

9.1. Identifying Sources of Endogeneity

Number of problems is encompassed by the term endogeneity. We can identify the different sources of endogeneity originating from the astronomical data. This can create problems if not addressed. Each object in universe might exhibit some unique features that are not shared by others. When the amount of data to be extracted is small, data points from small sub-populations are generally categorized so therefore it becomes hard to systematically model it due to insufficient observations [1].

$X_{it} = \mathit{planet\ rotation}$

Y_{it}= behavioral evaluation of objects.

data generating processes for these variables are

$$\gamma_{it} = \alpha_1 \gamma_{u-1} + \beta_{11}\, x_{it} + \sum_{m=2}^{M} \beta_{m1 z_{mit}} + u_{1it} \quad \text{......eq1}$$

$$x_{it} = \alpha_2 x_{it-1} + \beta_{12} \gamma_{it-k} + \sum_{l=2}^{L} \beta_{12 z_{lit}} + u_{2it} \quad \text{..eq2}$$

$$\mu_{jit} = \tau_t + \eta_{ji} + \epsilon_{jit} \quad \text{............. eq 3}$$

where Γ_t represents a common trend across all cases, η_{Ji} represents time invariant unobserved universal objects heterogeneity, and ϵ_{jit} are independent, identically, and normally distributed (3). The data-generating process suggests that x_{it} is a function of y_{it-k} (2). The z_{mit} are exogenous predictors of y_{it}, and the y_{it} exogenous predictors of x_{it} Generally, *k* in (2) is assumed to be either 0 or 1. This means that behavioral evaluations are a function of current planet rotation and/or a lag of planet rotation affect. It also considers least squares regression to estimate.

$$y_{it} = \beta_{01} + \beta_{11} x_{it} + v_{1it} \quad \text{.......eq4}$$

for T=1.ever other exogenous predictors are included,

$$v_{1it} = u_{1it} + \alpha_1 \gamma_{it-k}$$
$$= \eta_{1i} + \epsilon_{1it} + \alpha_1\, \gamma_{it-k} \quad \text{.........eq5}$$

As *T* = 1, this is a cross-sectional model and the exogeneity assumption for x_{it} for unbiasedness in the estimation of β_{11} is *E* ($v_{1it.}$, | $x_{it.}$)=0, and the exogeneity assumption for asymptotic unbiasedness as$\rightarrow$N .∞.is Cov($v_{it.}$, $x_{it.}$)=0 , If (1) and (2) do represent the data-generating process, the OLS estimation of the parameters in model (4) contains the following endogeneity problems. The first source of endogeneity is a consequence of not including the rotational effect of planets $\alpha_1\, y_{it-1}$, which predicts current values of both behaviour of star evaluations (when *k* = 1) and planet rotation in the model.

The different other sources of endogeneity includes (i) rapid increase of the data day by day (ii) no fix data/time varying data(iii)collection of massive of amount of data(iv)imagious data.(v)no control in the advancement in tools(v)difference in astronaut and public view point of data(vi)change in the pattern and the structure of representing data(vii)changing properties of astronomical objects .(viii)consequence of not including rotation and revolution effects (Xi)not controlling fixed properties of objects.

9.2. Potential Solution to Endogeneity

Huge amount of astronomical heterogeneous data implies both time varying data and to verify the similarity of stars within the clusters and to view the dissimilarity of stars among the clusters. We can prove this difference with the by using FFT coefficient within the stars[fig 3] , there is regressors that occurs in between it. To make difference more clear in this data generating process, equation 1 can be rewritten as

$$y_{it} = \alpha_1 y_{it-1} + \beta_{1,1.1} \overline{x}_i + \beta_{1,1.2}\left(x_{it} - \overline{x}_i \right) + \mu_{1it} \quad \text{...eq6}$$

The coefficient β_{111} captures the effects of average differences between objects (stars) as discussed in Figure 2, and β_{112} captures the effects of differences within stars over time. If (1) is estimated without making the distinction, the resulting estimate of β_{11} , will be an average of the two effects A common interpretation for the distinction is that the first is the effect of permanent (or at least long-term) differences in x_{it} , and the second is the effect of short-term temporal fluctuations in x_{it}. In this case, there may be long-term differences in behavioral evaluations across stars [24]. Different object in the universe bears some difference properties. These differences may be due to differences in properties of stars. The reasons for the differences are that the different objects belong to different classes and they are found at different part of universe. The properties of star ranges from difference in star light curves to the difference in the frequency domain.

However this paper will discuss two solutions. Solution 1 will be applied if we wish to verify the similarity of stars within the clusters. Solution 2 is to view the dissimilarity of stars among the clusters. From the survey we will start with solution 1 but before giving solution 2 we would consider the particular data set named as GALEX and STARS as described in the above sections.

$X_{it} = Revolution\ of\ planets$
Y_{it}= behavioural evaluation

9.3. Solution 1

With reference to Figure 2, we can see how the cluster with same kind of stars differs in the properties with that of the clusters with different kind of stars, this brings us to solution 1.

Now we can test effect of differences between the behaviour/properties of stars, we can test their properties when they are in the group with other stars i.e. for handing this kind of problem, the solution to the endogeneity problem is to instrument x_{it} . There are two important requirements for instrumental variables: (I)it should be partially correlated with x_{it}, that can control other exogenous variables in the model and secondly (II)it should not be correlated with the errors in the data, as they can include problem creating elements. The main concern for this approach is to find that measure that is independent of the errors in the model by being independent of y_{it-1},*and* η_{Ji} and more over it should be excluded with time-varying covariates that are included in the model. Because of this reason, the measure of behavioral evaluation was introduced that inquires about the exploration of properties of objects that is to be further used as a data. This measure can be considered as a part of indicator we have tested for their properties as the measure of individual object effecting the survey.

There are few disadvantages of solution I, that is sometimes there are possibilities that instruments are uncorrelated with error terms but in some cases still there are few chances that the instruments does always are like that. This kind of difficulty is mainly occurs in cross sectional data.

9.4. Solution 2

Testing for the effects of changes among the clusters(fig2,considering either LHS diag or only RHS diag) make use of panel data so that the estimation should be based on the main approach that is we are not interested in estimating incidental parameters rather we are interested in the estimation of common parameters that effects our data sets. So we find the re-parameterization of the incidental parameters in the way that incidental and common parameters are information-orthogonal. This let us to produce a estimation that is independent of the incidental parametric values. We can take this estimation as consistent → as N .∞. Therefore we continue to have incidental parameters but not an incidental parametric problem. The inclusion of $\alpha_1\ y_{it-1}$in (I) helps in removal of error terms there (from sol. 1). Hence we can resolve the problem (I). If we reparameterized the fixed effects, η_{Ji} so that they are information –orthogonal such that this allows us to estimate the interested parameters that of independent of other data. Through this way we can control the heterogeneity that was produced by these fixed effects and this resolves endogeneity problem 2. Omitted-variable problem can be reduced if we include additional dynamic covariates and hence it also exclude time varying covariates. And if we want to tackle with time varying data then we can make use of instrumental variables that have defined in the solution 1.

10. Data and Measurement

10.1. Extraction of Astronomical Big Numbers

As we are familiar about the fact that large numbers are found in astronomy. In this section we will discuss how the large numbers found in the surveys of astronomies are rounded off for the usage [12]. For example, according to the Big Bang model The whole Universe is 13.8 billion years (4.355×10^{17} seconds) old, and that the universe that is observable is 93 billion light years across (8.8×10^{26} metres), and it contains round about 5×10^{22} stars, which are organized into about 125 billion (1.25×10^{11}) galaxies, according to Hubble Space Telescope observations.

10.2. Combinatorial Process

Astronomical survey generates even larger numbers. Factorial numbers that are use to define the permutation on fixed set of objects is directly proportional to the number of objects, so in order to get the asymptotic expression for this rate of growth we make use of formula named as a Stirling's formula. Few numbers are too large that they are typically only referred by using their logarithms. For example, • 10^{10} (10,000,000,000), , called "ten billion" in the short scale and "ten milliard" in case of long scale. Even if the number is too large than we can make use of Conway chained arrow notation [12]. We can make use of the length of the chain.

10.3. Scientific Notations

In astronomy large number appearance is common, because of this many scientists make use of the standard notations. This helps at the time of rounding off very large decimated number. They will make use of terms mantissa and powers. Also, if the number is, 3×10^{39} tons, it will be equivalent to 3 000 000 000 000 000 000 000 000 000 000 000 000 000 tons, that is a 3 followed by 39 zeros. This number will be written in a scientific notation as a negative power corresponds to a small number. For example, the number 1×10^{-3} is written as 0.001 in conventional notation. In general, $10^{-n} = 1/10 \times 1/10 \times 1/10 \times ...$ for n times, Since scientific notation relies on powers of ten so it is quite easy to convert a number into a scientific notations.

10.4. Handling Uncertainties for Round Off

As we know those measurements are never perfect and numbers are never exact. So, every measurement give rise to some **uncertainty** related with it. Scientific notation shows the easy way to express how precisely a number is known. We can make use of the following two rules for solving arithmetic problems with large data [13]

- When multiplying or dividing numbers with uncertainties, make sure that the answer has as many significant figures as the least precise of the original numbers.
- When adding or subtracting numbers with uncertainties, round the result to the last significant place of the original number with the greatest uncertainty.
- Shift the decimal point and change the exponent. This is how the rounding off the big data is done.
- Else for the string data extraction we can make use of certain techniques like Data Mining (DM), Knowledge Discovery in Data Bases (KDD). They can make use of classification, clustering, catalogs for the purpose of the knowledge extraction.

11. Citation and Download of the Big Astronomical Data based upon Individual Needs

These differences may be due to differences in data need and data capturing time. People with different professions require data that is relevant according to their need. They don’t really need massive data, for example the researcher working in a NASA would require vast knowledge hence would expect data with deep meaning where as a doctor or a teacher teaching science as a subject would require only limited and to the point data. The other reasons for establishment of differences are time varying data. Day by day new research and sky surveys are being captured which can create chaos. The differences can be due to individuals with the different interests and they belong to different professions so they have different data needs.

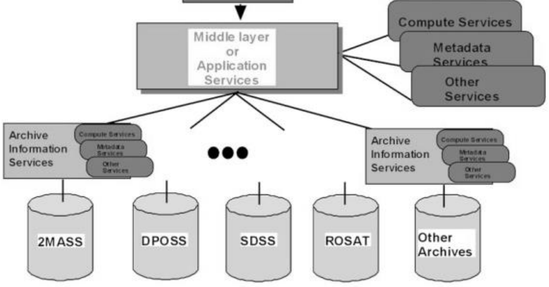
11.1. Format and Size

All astronomers ultimately think FITS (Flexible Image Transport System) as their data format for all their data needs. As per the data set size, the range is very large, with some small datasets, *e.g.*, in the range of few Megabytes for quasar flux density data have some medium-sized datasets. Such a platform allows citation and download of data. It also has the ability to select and download only a subset of the data available for a specific project, instead of entire dataset *e.g.*, a user should be able to select a region of sky delimited by coordinates (Right Ascension, Declination and an angular radius) and download matching observations for that particular region only.

11.2. The Dataverse Network

The Dataverse Network is an open source software application, The Dataverse enables the quick discoverability by searching across all in addition to information extracted from data files. The metadata is also mapped with various standard metadata schemas like that of Dublin Core and DDI and exported to XML format for preservation purposes .The Dataverse Network allows good practices for big scientific data publication: 1) supports metadata standards 2) it enables the inclusion of accompanying code and other materials for each and every dataset 3) it also provides the versioning of a dataset, with easy access to previous versions of the data and metadata, 3) assigns a persistent identifier (DOI) and generates a full data citation, with attribution to data authors and distributors 4) deep search for FITS files, helps in the indexing of the FITS files header information to facilitate discovery of such files 5)provide a data citation for every dataset uploaded. The citation includes a persistent identifier which links to data, and can be added to the references sections of any publication. The table can be added from published links. This helps the different types of users to grab data of their own interest and of their own use too.

11.3. Communication Fundamentals



The first requirement for connecting highly distributed datasets is that they should be able to communicate with each other. The communication performs multiple roles like discovering the holdings and capabilities of each archive, including the initiation of communication, the actual process of querying, the streaming of data, and overall control of structures. The language for communicating would be the extensible markup language (XML), using some standard schema. This will allow for control of the inter archive communication and processing *e.g.*, we can very easily perform the checkpoint operations on some query: pause, restart, abort, stop and at the same time we can provide appropriate feedback to the end users.

11.4. Performance Improving

Traditionally, astronomers were communicating with the data in the ASCII or by (FITS). The true efficacy of the FITS format is that it supports the streaming format, however there can be some difficulties like that of randomly extracting of desired data or shutting off the stream. The ideal solution is to pass different types of data (*i.e.* tabular, spectral, or imaging data) in a streaming fashion (similar to MPI—Message Passing Interface), so that there is no need for analysis of data to wait for the entire dataset before proceeding. In web services model, it will allow different services to cooperate in a head to tail fashion *i.e.* the UNIX pipe. This is a potential concern, because the ability to handle with XML encoded binary data is yet known.

12. Results

12.1. Comparing Solutions

If we compare both the solutions (sol 1, sol 2) then we can say that different properties of stars will definitely effect the data. Both current and changed values of the statistics (data variables) are use to instrument the data variable.

Current and changed value means that immediate change in the survey when two different types of objects mixes hence we can point out the change in the data **for solution 1-**The common approach between the properties of universal objects estimation is to run the cross-sectional analysis by using the average of cross-wave for each variable. Hence this helps in noise reduction from different reaction of individual star (fig3) **for solution 2-**As we have proved through solution 1 that change instrument is not likely to be endogenous. Change in individual object is represented by the instrument, so we can say that individuals object is also now not likely to be endogenous.

12.2. Experience, Lessons, and Observations

Based on the case study, there are limitations in existing technologies to support system performance and scalability evaluation. Here summarized experience, lessons, and observations are provided. It is clear that there is a need to have some ways through which we can extract the knowledge from the big data.

- **Astronomical KDD**

In the broadest sense, KDD/DM means the discovery of “models” for data. DM is one of the complex processes. In most cases the optimal results can be found only on trial and error bases by comparing the outputs of different methods or of different implementations of the same method. This implies that in order to solve a specific problem, we require a lengthy fine-tuning phase. Such complexity is among the reasons for a slow uptake of these methods by the community of potential users which still fail to adopt them to give the effective results, a DM application requires a good understanding of the mathematics underlying the methods, of the computing infrastructure, and of the complex workflows which are often needed to be implemented. So far, most domain experts in the scientific community do not make the effort needed to understand the fine details of the process, hence they prefer to recur the traditional approaches which are less powerful, but which may be more user-friendly.

- **Computational Requirements**

For maintaining the scalability, whenever there is a large quantity of data, the three main approaches are used to make it feasible. The first one is, applying of the training scheme to a decimated data set. But this can be very lengthy method and the information can be easily get lost [19]. The second one is splitting the problem in smaller parts (parallelization) sending them to different CPUs and finally combine the results together. The third one is to make the algorithms whose computational time is less [18].

12.3. Scientific Verification

There is a need to have a real time models for big astronomical data. Some of the scientific tests include-The new scientific approaches to star galaxy separation overcomes the restriction of current inaccuracies that effectively limits the scientific applications of any sky survey catalog. Related to this is an objective, automated, multiwavelength approach to morphological classification of galaxies, *e.g.*, quantitative typing along the Hubble sequence [19], or one of the more modern, multidimensional classification schemes. An automated search for rare and yet unknown astronomical classes of astronomical objects [21].

13. Conclusions

Beginning with the big data problem, reaching to its major challenges i.e. incidental endogeneity, dealing with the techniques and solutions by considering a certain astronomical data set, helps to address endogeneity problem in that data set and also gives two compact solutions for its removal after proving the result by using some standard formulas. Finally, discussing about extraction of such a big astronomical data.

Concluding, with the advancement in the technology like that of big telescopes (LSST) ,we can very easily get movie of whole sky which continuously give rise to big data and ultimately becomes the reason for endogeneity. So with development of technology there should be equal development for certain tools for scientific exploration. Hence it becomes essential to create a environment that will be purely interactive, dynamic and web-based for research in astronomy with huge data sets. In such an environment it would be easy to have deep research about our universe, which would help to know where the can life can exist. It can be helpful to create the medicines for newly rising diseases. Hence can saving many lives. This can take the India to the time where it would be called a Developed nation. This can be made possible if and only if the big data will be free from the endogeneity problem.

Acknowledgements

We are very grateful to our advisor Pankaj Deep Kaur for her support, patience, motivation and immense knowledge. We could not have imagined of having a better advisor than her. Her guidance helped us in all the time of writing this survey paper.

References

[1] J. C. Aubele, L. S. Crumpler, U. M. Fayyad, P. Smyth, M. C. Burl, and P. Perona, “Locating small volcanoes on venus using a scientisttrainable analysis system”, In Lunar and Planetary Science Conference, vol. 26, **(1995)**, p. 61.

[2] T. S. Axelrod, “Statistical Issues in the Macho Project", Mt Stromlo and Siding Spring Observatories, The Australian National University, **(1996)**.

[3] R. R. de Carvalho, S. G. Djorgovski, N. Weir, U. Fayyad, K. Cherkauer, J. Roden and A. Gray, “Clustering analysis algorithms and their applications to digital possii catalogs”, In Astronomical Data Analysis Software and SystemsIV, (**1995**), vol. 77, of ASP Conference Series, pp. 272–275.

[4] R. J. Brunner, S. G. Djorgovski, and A. S. Szalay, “VirtualObservatories of the Future”, Astronomical Society of the Pacific, San Francisco, **(2001).**

[5] R. G. McMahon and M. J. Irwin, “Apm surveys for high redshift quasars”, In ASSL Vol. 174: Digitised Optical Sky Surveys, **(1992),** p. 417.

[6] R. Duda and P. Hart, “Pattern Classification and Scene Analysis”, John Wiley, & Sons (**1981**).

[7] L. M. Barrosaro, “Data Mining in Large Astronomical Databases”, To appear in Astrostatistic Springer Series on Astrostatistics **(2011).**

[8] http://en.wikipedia.org/wiki/Large_numbers

[9] http://en.wikibooks.org/wiki/General_Astronomy/Print_version

[10] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” in Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation (OSDI**),** vol. 6, **(2004)**, (Berkeley, CA, USA), pp. 10–10, USENIX Association.

[11] C. T. Chu, “Map-reduce for machine learning on multicore”, in NIPS (B. Sch¨olkopf, J. C. Platt, and T. Hoffman, eds.), pp. 281–288, MIT Press, **(2006)**.

[12] Fan and Li **(2001)**, Hunter and Li **(2005)**, Zou **(2006)**, Zhao and Yu **(2006)**, Huang, Horowitz and Ma **(2008)**, Zhang and Huang **(2008)**, Wasserman and Roeder (2009), Lv andFan **(2009),** Zou and Zhang **(2009)**, Städler, Bühlmann and van de Geer **(2010)**, and Bühlmann, Kalisch and Maathuis **(2010)**.

[13] X. Chen and D. Pouzo, “Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals”, Econometrica, vol. 80, **(2012)**, pp. 277–321.

[14] S. A. Van de Geer, “High-dimensional generalized linear models and the lasso”, **(2008),** *Ann. Statist*.**36** 614–645. MR2396809.

[15] R. E. Rutledge, R. J. Brunner, T. A. Prince and Lonsdale, “CXID: Cross Association of ROSAT/Bright Source Catalog XRay Sources with USNO A2 Optical Point Sources”, *Astrophysical Journal, Supplement*, **(2000).**

[16] Sources with USNO A2 Optical Point Sources, *Astrophysical Journal*, *Supplement*, **(2000)**.

[17] A. Szalay and R. Brunner, “Astronomical Archives of the Future: A Virtual Observatory”, *Future Generations of Computational Systems*”, **(1999).**

[18] R. G. McMahon and M. J. Irwin, Apm “Surveys for high redshift quasars”, In *ASSL Vol. 174: Digitised Optical Sky Surveys*, **(1992)**, p. 417.

[19] S. C. Odewahn, E. B. Stockwell, R. L. Pennington, R. M. Humphreys, and W. A. Zumach, “Automated star/galaxy discrimination with neural networks”, *Astronomical Journal*, **(1992).**

[20] C. C. Chang and C. J. Lin, “*Training Support Vector Classifiers: Theory and Algorithms*, Neural Computation”, **(2001).**

[21] E. Meng Joo and L. Fan, “*Genetic algorithms for MLP neural network parameters optimization,* in Proc. 21st Intl. Control and Decision Conference, Guilin, China”, IEEE Press, (**2009).**

[22] J. D. Angrist and A. B. Krueger, “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments”, *Journal of Economic Perspectives*, **(2001).**

[23] C. J. Anderson, “The End of Economic Voting? Contingency Dilemmas and the Limits of Democratic Accountability”*, Annual Review of Political Science, (***2007)**.

[24] C. J. Anderson, S. M. Mendes, Y. V. Tverdova, and H. Kim, “Endogenous Economic Voting: Evidence from the 1997 British Election”, *Electoral Studies, (***2004)**.

Authors

Sumedha Arora, done B-tech from Khalsa College of Engineering and Technology and pursuing, Mtech (11nd) year from GNDU.

Pankaj Deep Kaur, he is an Assistant professor at GNDU (RC-Jalandhar).