# Generalized Numerical Lattices for Time Series Representation in Complex Data Systems

**Thalita B. Veronese**[1], **Reinaldo R. Rosa**[1], **Nandamudi L. Vijaykumar**[1], **Mauricio J.A. Bolzan** [2]

[1]Núcleo para Simulação e Análise de Sistemas Complexos
Laboratório Associado de Computação e Matemática Aplicada (LAC)
Instituto Nacional de Pesquisas Espaciais (INPE)
12201-970 São José dos Campos, SP, Brazil

[2]Instituto de Pesquisa e Desenvolvimento
Universidade do Vale do Paraíba (UNIVAP)
São José dos Campos, SP, Brazil

{thalita,reinaldo,vijay}@lac.inpe.br, bolzan@univap.br

***Abstract.*** *Analysis of information from multiple data sources obtained through high resolution instrumental measurements has become a fundamental task in all scientific areas. The development of expert methods able to treat such multi-source data systems, with both large variability and measurement extension, is a key for studying complex scientific phenomena, especially those related to systemic analysis in space and environmental sciences. In this paper, we propose a times series generalization introducing the concept of generalized numerical lattice, which represents a discrete sequence of temporal measures for a given variable. As a case study, we show a preliminary application in space science data, highlighting the possibility of a real time analysis expert system to be developed in a future work.*

## 1. Introduction

Scientific research based on high resolution measurement instruments (sometimes combined with high resolution numerical simulations) leads to systemic multiple data sources resulting in heterogeneous scientific data, that we call *complex data systems*. As a consequence, modern science is confronted with a large variety of high resolution data, advanced mathematical techniques and algorithms for multidimensional data analysis. Thefore, a development of importance to scientific computing in general is the representation of the enormous amount of information as organized and coherent database systems. A special attention should be devoted to complex data system composed of time series generated from *complex systems* observation. New concepts such as *complex systems* are related to real systems in physics, chemistry, biology, economics, etc, when they are characterized by collective, time-dependent phenomena emerging from the dynamic interplay of a large number of heterogeneous constituents, observed and analyzed in detail [P.E.Cladis and P.Palffy-Muhoray 1995, Yanner-Bar-Yan 1997]. It means that there will be observations of structures and processes in all possible temporal, spatial and spectral scales. Because of this multi-scaling and multidimensional approach, information on nonlinearities, long-range correlations and phase transitions should be present in a post-analytical data representation. For this reason, in this paper, we introduce an innovative

generalization which represents a set of time series based on: (i) its constitutive variables $U_i(t, s)$ as function of time ($t$) and euclidian space ($s(x,y,z)$); (ii) its time and space discrete extensions; and (iii) second-order measurements coming from the analysis of each $U_i(t, s)$. Hence, in our approach the concept of *generalized numerical lattice* (GNL) is introduced based on structural and phenomenological information of a discrete sequence of spatio-temporal measures for a given constitutive variable. This paper presents a case study introducing the GNL concept to an integrated data representation for the Brazilian space weather program. In this program there are several data files coming from high resolution observations taken by different satellites and ground instruments. There are data from the solar atmosphere, interplanetary medium, magnetosphere and ionosphere observed in one (1D), two (2D) and three (3D) spatial dimensions with different time and space resolution and different frequencies, all representing distinct physical processes possibly nonlinearly correlated. Thus, the main goal in a space weather program is to follow the Sun-Earth magnetic coupling to understand the solar-terrestrial relationship and predict geoeffective events. Furthermore, it will be clear from the following sections that GNL-based data representation can be useful for data mining, advanced data analysis and information representation systems in many different scientific areas where a robust data description from a complex data system is required.

## 2. Complex Data Systems

In most scientific areas, there are many ways to collect data from natural systems in order to extract data structural information and perform different kinds of analysis on them. For this reason, researchers usually have a lot of data sets stored independently, occupying huge hard drive memory space, which increases with the technological advances. In this context, data systems are often composed by spatio-temporal information of 1D, 2D and 3D which can represent many distinct possible measurements taken from the same observed system.

Nowadays, for example, a data system from a solar active region is composed of many time series observed in almost all electromagnetic spectra (radio, visible, infrared, UV, x-ray) in 1D and 2D, plus a set of possible correlated data from the interplanetary field, magnetosphere and ionosphere [SPIDR ]. Also, data from numerical simulation based on magnetohydrodynamic (MHD) models can be addressed [Buchner et al. 2003]. Thus, space weather investigation, for example, is a promising application where more than fifty different kinds of time series data are available involving observations and simulations in all possible spatio-temporal physical dimensions [Hanslmeier 2007]. Generally, space physics data are generated from international space programs administered by [NASA ], [NOAA ] and [ESA ].

As illustrated in Figure 1, the task of modeling the data system in an organized and meaningful representation is achieved by the execution of a sequence of dependent steps. First, real systems observations are performed using several instruments and/or numerical experiments. The resulting measurements can be organized as metadata and data files. Then, data processing is performed, which refers to a class of programs that organize and manipulate data, usually large amounts of numeric data. Next, the data are usually visualized and analyzed. A post-analytical data acquisition system is a device designed to measure and log some data parameters. In a nutshell, one can acquire measurements over the observable system in as many variables as desired, then perform a parameterization

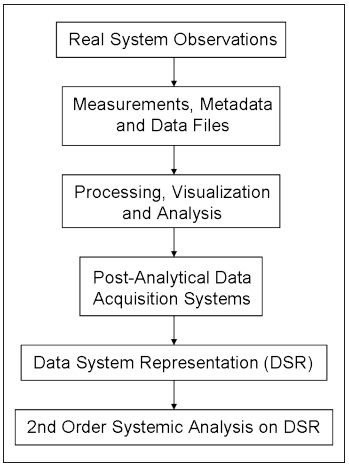


**Figure 1. From the real system observation to an useful data system representation.**

and analyze the information in order to put them in the numerical lattice representation, as proposed in next section. The final model is obtained by gathering all files – one for each kind of data – in a single *Data System Representation* (DSR), from which systemic analysis might be performed to identify, for example, space weather features responsible for geostationary satellite anomalies.

In this paper we introduce the concept of Generalized Numerical Lattice (GNL) in order to generate a post-analytical data integration which we call here a *GNL-based Data System Representation*.

## 3. Generalized Numerical Lattices

Numerical lattices are interpreted here as any regular and discrete distribution of numerical quantities structured in a Cartesian space bounded by linear spatial dimensions. We propose here a new formalism for a mathematical generalization of data systems obtained from multiple measures over a single system, based on the concept of *generalized numerical lattices* (GNL) which represents any dynamical sequence of measurements of a constitutive variable $U_i(t,s)$. A GNL is defined as a structural data representation, $\pounds$, where a given time series is represented by three kinds of coefficients, being the first and second for the data structure (function and extension domains, respectively) and the third kind for post-analytical properties as statistical moments, power spectrum index, morphometrical quantities, etc. Thus $\pounds$ can be written as

$$\pounds = f\left(\kappa, \lambda_\ell, \mu_p\right) \quad , \tag{1}$$

where $\kappa$ is defined as being the *variational degree*, which is the amount of state variables from the fundamental domains (time and three-dimensional Euclidian space); $\lambda_\ell$ indicates the *extension coefficients*, given by the quantity of discrete measures at each usual domain; and $\mu_p$ is the set of *post-analytical properties* characterizing the dynamics and/or statistical behavior of the constitutive variable $U_i(t,s)$.

### 3.1. The Variational Degree: $\kappa$

The variational degree $\kappa$ depends on how many of the possible kinds of constitutive variables from the real system are available in the data system. In a GNL all variables that can be measured are considered varying at least in time. As shown in Table 1, time ($t$) is always present and is characterized by $\kappa = 1$. So we can have, for a 1D discrete sequence of data, one temporal observable $U(t)$, correspondng to $\kappa = 2$. If we have space-time information measured in one, two or three dimensions, respectively, the variational degree increases to 3, 4 or 5. For a spatio-temporal series (e.g., a sequence of images composed by $\lambda_1 \times \lambda_2$ pixels) we have $\kappa = 4$. When the data is a dynamical hypercube composed by $\lambda_1 \times \lambda_2 \times \lambda_3$ voxels the GNL has $\kappa = 5$. Note that GNL composed by extra dimensions and new coupled variables (functionals) correspond to $\kappa \geq 6$.

### 3.2. The Extension Coefficients $\lambda_\ell$

Extension coefficients $\lambda_\ell$ refer to data set length in each measured variable. So, $\lambda_0$ refers to the number of points $N$ that compose the vector $U(t)$; $\lambda_1$ is the size of the data in the $x$ Euclidian spatial domain; $\lambda_2$ is the size of the next discrete dimension $y$ , and so on. Thus, a $\pounds_{2,\lambda_0}$ represents a $U(t)$ time series composed by $\lambda_0$ points, while a $\pounds_{4,\lambda_0,\lambda_1,\lambda_2}$ represents a spatio-temporal series composed by $\lambda_0$ images of size $\lambda_1 \times \lambda_2$ with a intensity measure $U(t, x, y)$ in each correspondent pixel $(x, y)$. As examples, a $\pounds_{2,10^4}$ represents a $U(t)$ time series composed by 10.000 points, while a $\pounds_{4,10^2,64,64}$ represents a dynamical sequence of 100 images of size $64 \times 64$. A given GNL $\pounds_{5,10^2,64,64,64}$ represents a dynamical sequence of 100 hypercubes of size $64 \times 64 \times 64$ with a intensity measure $U(t, x, y, z)$ in each correspondent voxel $(x, y, z)$.

**Table 1. The constitutive variables as a function of the variational degree $\kappa$.**

| $\kappa$ | Constitutive Variables |
|---|---|
| 1 | $U_1 = t$ |
| 2 | $U_1 = t; U_2 = f(t)$ |
| 3 | $U_1 = t; U_2 = x; U_3 = f(x, t)$ |
| 4 | $U_1 = t; U_2 = x; U_3 = y; U_4 = f(x, y, t)$ |
| 5 | $U_1 = t; U_2 = x; U_3 = y; U_4 = z; U_5 = f(x, y, z, t)$ |
| 6 | $U_1 = t; U_2 = x; U_3 = y; U_4 = z; U_5 = g(U_i, i \leq 4); U_6 = f(x, y, z, U_5, t)$ |
| $\vdots$ | $\vdots$ |

### 3.3. The Post-analytical Parameters $\mu_p$

There are several post-analytical parameters $\mu_p$ which are relevant for time series characterization [Dunn 2005, Peitgen et al. ]. When $\kappa = 2$, the autocorrelation of $U(t)$ is the first property to be considered. Thus, $\mu_1$ is the cross-correlation of $U(t)$ with itself, a measure with values in the interval: $-1 \leq \mu_1 \leq 1$, which characterizes repeating intensities, such as the presence of a periodic signal which has been buried under noise, or the missing fundamental frequency in $U(t)$ imposed by its harmonic frequencies [Dunn 2005]. When the variability pattern of $U(t)$ is a perfect Gaussian white noise we have $\mu_1 \approx 0$ (non-correlated with a normal probability distribution). Hence, $\mu_1$ is able to detect non-randomness in data and can be used to identify an appropriate time series model when $U(t)$ has a deterministic component. Consequently, a second kind of useful

post-analytical property is the characterization of $1/f^{\mu_2}$ noise from the power-spectrum of $U(t)$ [Keshner 1982]. Here, the power-law spectral index $\mu_2$ is used to identify the scales in which the lattice presents stronger correlation. The correlation level can be formulated either in simple frequency form or in cumulative frequency form, usually as a rank-size type relationship, which is preferred in this case, when the focus is on the rarer or larger events that dominate the distribution of $U(t)$ for different temporal scales. There are many other post-analytical properties as Kullback-Leibler divergence [Burnham and R. 2002], fractal-like dimensions, Kolmogorov-Sinai entropy, Hurst exponent [Peitgen et al. ], singularity spectral index [bol ], etc, which can be addressed for GNL with $\kappa = 2$ (1D time-series case $U(t)$). Although the quantity of post-analytical measures is an open set $\{\mu_1, ..., \mu_j, ..., \mu_p\}$, here we are considering $\mu_1$ (the autocorrelation coefficient) and $\mu_2$ (the power-spectrum index) for $\pounds(2, \lambda_\ell, \mu_1, \mu_2)$. However, when $\kappa = 4$, a third property is given by morphometrical and/or image processing measures. Some examples of post-analytical properties for the case $\kappa = 4$ are spatial correlation functions, Minkowski functionals [K.R.Mecke and D.Stoyan 2000] and *gradient moments* from the Gradient Pattern Analysis (GPA) [Rosa et al. 1999, Rosa et al. 2003, R.R.Rosa et al. 2007], which characterize 2D Physical information of the $U(t, x, y)$ pattern observed in the spatio-temporal domain (see Table 1).

Without loss of generality, in this paper we use the following final notation for a given GNL: $\pounds_{\kappa,\lambda_0,...,\lambda_\ell}(\mu_1, \mu_2)$, with $\mu_1$ and $\mu_2$ representing, respectively, the auto-correlation coefficient and the power-spectrum index, which can be calculated for $U(t)$, $U(t, x, y)$ and $U(t, x, y, z)$. Such parameters are used here as the simplest examples for $\mu_1$ and $\mu_2$.

Taking the examples given in Section 2.2, a $\pounds_{2,10^4}$ (0.28,-1.66), hence, represents a $U(t)$ time series composed by 10.000 points, with auto-correlation equals to 0.28 and power-spectrum index equals to -1.66. Such values are revealed to diagnose turbulent-like behavior and, hence, can suggest a process or modeling for the $U(t)$ time series. In this example, $(\mu_1, \mu_2)$ are post-analytical properties explicitly represented in a given Generalized Numerical Lattice. Examples for the cases when $\kappa = 4$ and $\kappa = 5$ can be easily perceived.

### 3.4. A Data System Representation

Taking into account a set of generalized numerical lattices representing a collection of experimental measurements of a given observed system, a post-analytical Data System Representation (DSR) consists of a grid containing all the GNLs relative to a particular data system, with the lines arranged by $\lambda_0$ (the data temporal extension) and the columns, by $\kappa$, both in ascending order, as in the example illustrated in Figure 2. The post-analytical parameters $(\mu_1, \mu_2, \ldots, \mu_p)$ are located under each respective GNL. An important property of this DSR is that the right-hand column and the bottom row are marginal totals. The right-hand column gives the marginal total for GNL with the same $\kappa$ and the bottom row gives the marginal total for GNL (the values are organized following $\lambda_0$ increasing), so that the box in the bottom right-hand corner is the grand total $L$ of GNL considered. Note that the DSR compute and show the total amount of GNL representing the data system. When the data system is represented for all values of $\kappa$, the right-hand column gives the marginal total of time series $U(t)$ (1st line), $U(t, x)$ (2nd line), $U(t, x, y)$ (3rd line), $U(t, x, y, z)$ (4th line), etc. When there is no data for some $\kappa$, the representation takes the

next $\kappa$ automatically. For example, when there is no $\pounds_3$ ($U(t,x)$), the total of time series $\pounds_4$ ($U(t,x,y)$) is shown in the right-hand of the second line.

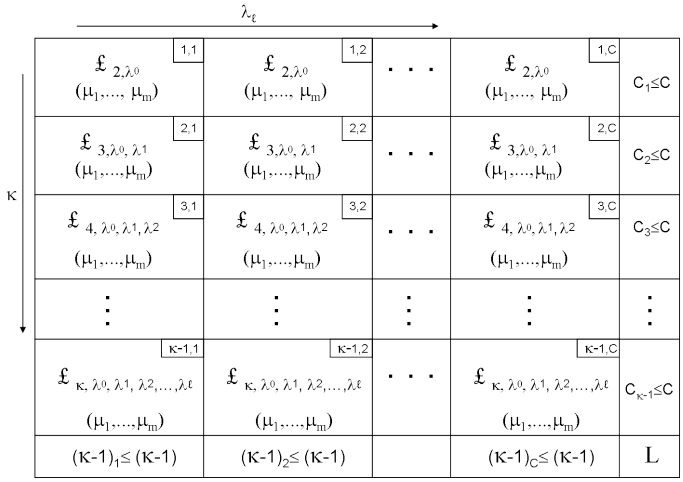


**Figure 2. The GNL-based Data System Representation.**

This GNL-based DSR may be helpful in many areas and applications, including Data Mining based on structural properties of time series, multidimensional data modeling, multivariate informations systems, specially those obtained in space and environmental physics, genomics, neuroscience besides other spatio-temporal databases in general sciences.

## 4. A Case Study in Space Physics

Space science data often need to be systemically analyzed in order to obtain mutual information necessary for space weather forecasting. Usually, the analysis is necessary to understand the physical processes studied by identifying and understanding the interrelationships of different parameters. In other cases it is necessary to use the data to build a model of the solar process which are geoeffective. In this context, solar activity is one of the main sources of space disturbances, which are primarily responsible for space weather phenomena observed in the Interplanetary Medium and in Earth plasma atmosphere.

In this section we show an example of using *GNL-based DSR* for multiple data sources based on the Space Weather Program which has been developed at [INPE ]. Initially, we have performed a preliminary data selection from three data sources: (i) The Space Physics Interactive Data Resource [SPIDR ], (ii) The NASA International Solar-Terrestrial Physics [ISTP ] and (iii) The Ondrejov Solar Radio Data Archive [ASU ]. The metadata we have selected to illustrate this application are shown in next section, all related to the space weather activity observed from June 5-8, 2000.

### 4.1. The Simplest GNL-based DSR

Table 3 shows the GNL-based DSR for the June 6, 2000 Solar Activity (SAJ6). The GNL are obtained from the data shown in Table 2. The quantities inside the parenthesis show the post-analytical properties obtained over these data. Note that, when some property is missing, it is indicated in the lattice keeping the correspondent identification $\mu_p$.

**Table 2. Selected data for SAJ6.**

| Data | Instrument (Source) | Temporal Size ($N = \lambda_0$) | Spatial Size ($\lambda_\ell, \ell > 0$) |
|---|---|---|---|
| Radio Burst 3GHz | Ondrejov (Sun) | 2988 | - |
| X-Ray Flux | GOES (Sun) | 5760 | - |
| Ion Density | ACE (IMF) | 5600 | - |
| Dst | Kyoto (MAG) | 8736 | - |
| UV Image | TRACE (Sun) | 3 | 256 |
| WL Image | TRACE (Sun) | 9 | 512 |

**Table 3. The Simplest GNL-based DSR for SAJ6**

| | | | | |
|---|---|---|---|---|
| $\pounds_{2,2988}$ $(0.6, -1.69, \mu_3)$ | $\pounds_{2,5600}$ $(0.4, -1.87, \mu_3)$ | $\pounds_{2,5760}$ $(0.5, -1.56, \mu_3)$ | $\pounds_{2,8736}$ $(0.3, -1.28, \mu_3)$ | 4 |
| $\pounds_{4,3,256,256}$ $(\langle 0.8 \rangle, \mu_2, 1.92)$ | $\pounds_{4,9,512,512}$ $(\langle 0.6 \rangle, \mu_2, 1.96)$ | | | 2 |
| $\pounds_{5,1,64,64,64}$ $(\mu_1, \mu_2, \mu_3)$ | | | | 1 |
| 3 | 2 | 1 | 1 | 7 |

A more complete GNL-based DSR for SAJ6 is under construction, using all available data from SPIDR and ISTP. In our complementary work, using the future expert system for GNL-based DSR, we will take into consideration also data from the space programs administered by the Brazilian National Institute for Space Research [INPE ].

## 5. Concluding Remarks

Here we propose a new concept for physically meaningful data generalization, that we name generalized numerical lattices (GNL). For a given data system, every time series which can be interpreted as a GNL is organized in a GNL-based Data System Representation. We expect to provide with this new data representation a better way for modeling and understanding complex data systems, avoiding redundant analysis and storing of information, and, moreover, introducing new methods of systemic characterization. The GNL-based DSR algorithm is still in progress, and implementations may be done soon to confirm its scientific applicability. To be effective, it must be simple enough to build the GNL-based DSR structure to the end user, supporting a 2nd order systemic analysis, which, due to the present scientific purpose, was not developed in this paper.

Besides the space physics data representation exemplified in this paper, the GNL-based DSR may also be helpful in many areas and applications, including Data Mining based on phenomenological properties in physics, chemistry and biology. Special attention should be paid to multidimensional time series data modeling and multivariate analysis in Environmental Physics, GIS, Neuroscience, Genomics and Astrophysics.

## Acknowledgment

The authors would like to thank CNPq for partial financial support. The authors are grateful to data from SPIDR and NASA, and to Ramon Morais de Freitas for interesting

discussions on GNL. The Transition Region and Coronal Explorer, TRACE, is a mission of the Stanford-Lockheed Institute for Space Research, and part of the NASA Small Explorer program. Yohkoh data are provided by NASA/ISAS.

## References

ASU. Ondrejov solar radio event archive info. http://www.asu.cas.cz/r̃adio/info.htm.

Buchner, J., Dum, C., and Scholer, M. (2003). *Space Plasma Simulation*. Springer.

Burnham, K. P. and R., A. D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Dunn, P. F. (2005). *Measurement and Data Analysis for Engineering and Science*. McGraw-Hill.

ESA. European space agency. http://www.esa.int/esaCP/index.html.

Hanslmeier, A. (2007). The sun and space weather. 347.

INPE. National institute for space research. http://www.inpe.br.

ISTP. http://istp.gsfc.nasa.gov/istp/events/2000jun6/.

Keshner, M. (1982). 1/f noise. *Proceedings of the IEEE*, 70(3):212–218.

K.R.Mecke and D.Stoyan (2000). *Statistical Physics and Spatial Statistics*. Springer-Verlag.

NASA. National aeronautics and space administration. http://www.nasa.gov.

NOAA. National oceanic and atmospheric administration. http://www.noaa.gov.

P.E.Cladis and P.Palffy-Muhoray (1995). *Spatio-Temporal Patterns in Nonequilibrium Systems*. Addison-Wesley.

Peitgen, H.-O., Jürgens, H., and Saupe, D. *Chaos and Fractals: New Frontiers of Science*. Springer.

Rosa, R. R., Campos, M. R., Ramos, F. M., Fujiwara, S., and Sato, T. (2003). Gradient pattern analysis of structural dynamics: application to molecular system relaxation. *Braz. J. Phys.*, 33:605–609.

Rosa, R. R., Sharma, A. S., and Valdivia, J. A. (1999). Characterization of asymmetric fragmentation patterns in spatially extended systems. *Int. J. Mod. Phys.*, 10:147–163.

R.R.Rosa, M.P.M.A.Baroni, G.T.Zaniboni, da Silva, A., L.S.Roman, J.Pontes, and M.J.A.Bolzan (2007). Structural complexity of disordered surfaces: Analyzing the porous silicon sfm patterns. *Physica A*, 386(2):666–673.

SPIDR. Space physics interactive data resource. http://spidr.ngdc.noaa.gov/spidr/home.do.

Yanner-Bar-Yan (1997). *Dynamics of Complex Systems*. Westview Press.