USER MANUAL
AnalisisSerial

Junio 18, 2010
Last revision: September 2019
Ruben Chaer
Montevideo - Uruguay.

1. Introduction.

The AnalisisSerial program is an auxiliary utility for the SimSEE platform. AnalisisSerial is useful for analyzing time series of data and creating a Gaussian Space Correlations model with CEGH Histogram.

2. CEGH Model.

Given a set of time series, with uniform synchronized sampling between the series that cover the same time horizon, it is possible to identify a model that represents the series set, maintaining some important characteristics of them.

The first question is, what is the model for?

The first answer is that the model is used in SimSEE to generate synthetic time series with the same statistical characteristics as the set of data series used to create the model.

In addition to the simple possibility of generating synthetic time series with characteristics similar to the historical series, the model attempts to capture the structure of the random process, creating a representation of The State of the system. The State of the system is, by definition, the vector of information that captures the relevance of the past of a system. Knowing the State, it is possible to calculate the future evolution of the system if the values of its future entries are known. This characteristic of State model is what makes it possible to consider the stochastic process in the Stochastic Dynamic Programming algorithms. In other words, it is what allows SimSEE to generate system operation policies that take into account the status of stochastic processes. As an example, if knowledge of the surface temperature of the Pacific Ocean in the area known as N34 has an influence on the probabilities of rainfall in the following months, knowledge of that variable determines the probabilities of the available hydroelectric energy in the following months and therefore, determines what will be the optimal water use policy for reservoirs.

In the stochastic dynamic programming, the calculation of the future cost function (or Bellman's function) on the state space is performed recursively (Bellman's recursion) from the FUTURE to the PRESENT (inverse time), so a coherent way of generating the random series values (for example the contribution to the dams) from the state of the system is necessary. In short, if the series set has no status (that is, it has no memory), it would not be necessary to identify a model that captures that memory to represent it in the stochastic dynamic programming algorithm. But if the memory of the hidden system that generates the data series matters, it will be necessary to have a model that represents that memory in order to carry out optimizations that depend on the state.

In other words, when "THE STATE" of the system is represented, the State of the stochastic processes (whose status is relevant for the purposes of what is being studied) must be included, in order to be considered in the stochastic dynamic programming algorithm . If the State needs to be identified, we must have a hidden system model that generates the series. A possible model is the CEGH, another model can be a chain of Markov states.

Below is a summary of the CEGH model, enough to understand the result generated by AnalisisSerial. For a more detailed description of the fundamentals of CEGH models see .[1].

The idea behind the CEGH model is to achieve a model that is capable of generating series with the same histogram of amplitudes as the original series while maintaining cross-correlation functions between the series and the series with themselves and their pasts.

The key to CEGH models is in: 1) The probability density function of a Gaussian stochastic process is completely determined by the Auto-correlation function (for a process of several variables, the definition of self-correlation with the matrix of covariances is generalized) and 2) The function of auto-correlation of a signal is the Fourier anti-transform of the signal power spectrum.

Then, given a model that generates the same power spectrum as that of the historical series, there will be a model that generates series with the same auto-correlation functions (or covariance matrices for the multi-variable case) than the historical one.

If, in addition, the process is Gaussian, as the probability density functions are determined by the coefficients of the self-correlation function (or the covariance matrices for the multi-variable case), a model will be generated that generates series with the same functions of probability density and therefore able to give the same probability measures.

Given a time series, there is an important arsenal of tools to work in the frequency domain for the synthesis of a linear filter (or linear system) whose frequency response matches the signal power spectrum. That filter, when fed with white noise at its input, generates time series with the same auto-correlation function as the original series. In this way for a series we have a powerful tool to generate models that capture the memory of a system.

Another important result of linear systems is that the output of a linear filter that is fed with Gaussian white noise is Gaussian.

The CEGH model combines both results, creating a set of non-linear transformations (Deformers) that attempt to transform historical series into a Gaussian process (normalized with series of mean null value and unit variance). These Deformers are invertible and given a Gaussian series with normal distribution N (0,1) they can anti-transform it into series with the same histograms of amplitudes as the original series. In the Gaussian space (corresponding to the historical signals transformed to Gaussians), we proceed to identify a linear filter (or linear system) that generates time series that represent Gaussian processes with correlation and self-correlation functions that approximate the corresponding ones of the original series of

data with which the model was identified. Once the linear system is identified, it is used to generate synthetic series by feeding the system with Gaussian white noise. The outputs of the linear system are time series with the desired correlations and self-correlations but with the amplitudes corresponding to Gaussian variables. To obtain the histograms of original amplitudes, the inverse of the non-linear transformations identified in the model construction process are used. The series thus transformed, by construction, have the same histogram of amplitudes as the original series of data.

Fig.1 summarizes the above, where $rbg(t)$ is a vector of Gaussian white noise sources that attacks a linear filter generating the output vector $xg(t)$. The $xg(t)$ are the outputs of Gaussian processes with the correlations imposed by the linear filter. Then, the series $xg(t)$ are transformed by a set of non-linear transformations $TNL(x,t)$ thus obtaining the vector of synthetic series $y(t)$.

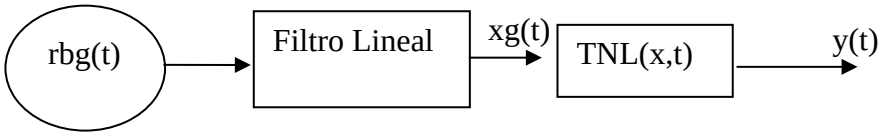

Fig. 1: CEGH model. Procedure of synthesis of values.

The AnalisisSerial program performs the analysis of a set of data series and as a result produces the "Linear Filter" and the TNL transformations, thus creating what is necessary to define the CEGH model of the process. The generated model is usable in SimSEE, both in the optimization stage and in the simulation.

3. Executing AnalisisSerial.

The AnalisisSerial program is distributed together with SimSEE.

If you use Windows, to run the program just go with the Windows explorer to the folder C: \ SimSEE \ bin and open the executable "AnalisisSerial.exe".

If you use Linux, to run the program use a terminal in your preferred graphical environment and in the SimSEE binaries directory {$ HOME} / SimSEE / bin run the "analisisserial" binary.

The screen that opens when executed is the one shown in Fig.2.

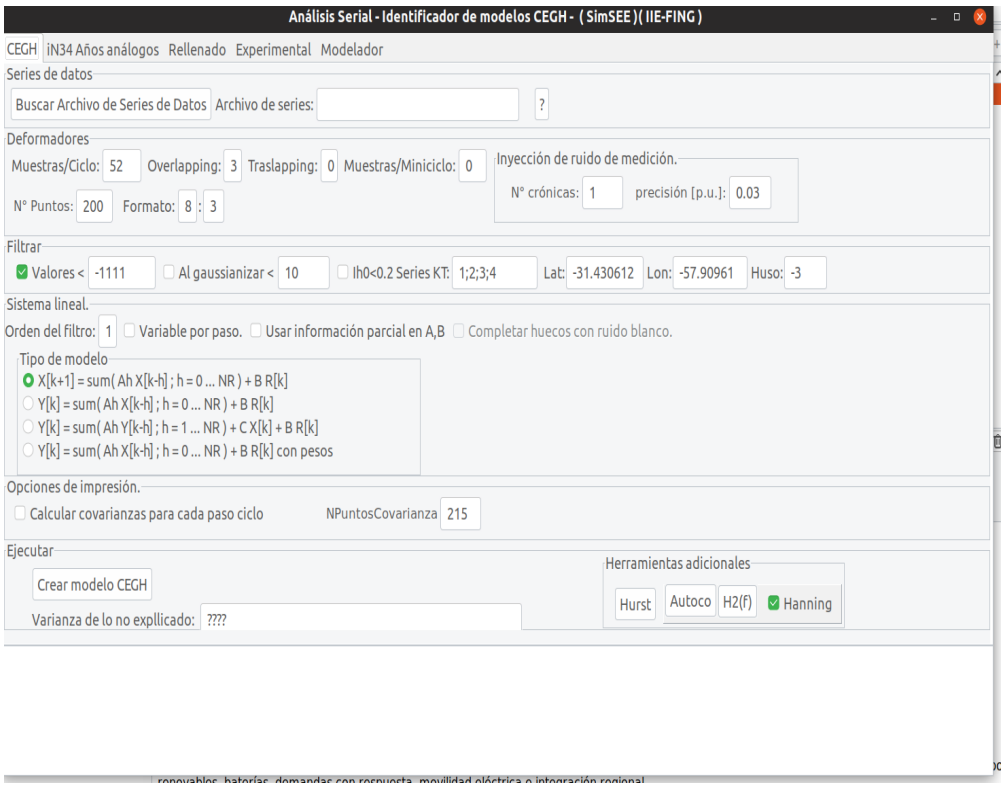

Fig. 2: AnalisisSerial Screen.

As you can see, at the top is the "Search Data Series File" button that allows you to select the file with the data series to process. This file must contain the data series with the format specified in the next section.

Then you must specify some parameters (Filter Order, Overlaping, etc.) and press the "Create CEGH Model" button to execute the analysis of the input data and write the results to the output file.

Once the calculation is finished, the box "Variance of the unexplained" is written, which is a measure of the unexplained, for each channel (data series), by the identified filter.

As an example, for a case of 2 data series, the output printed in the box corresponding to the Variance of the unexplained was:

[2|0.388388116871901;0.385542413429916]

The first 2 (followed by the "|" character) indicates that two numbers follow, one corresponding to each time series. The character ";" (semicolon) is used as the separator of the list of numbers. In the case of the example, the variance of the unexplained is 0.388388116871901 for the first channel, and 0.385542413429916 for the second.

If for a channel the variance of the unexplained was 1, it means that it was not possible to capture useful information from the input series, and therefore everything remains to be explained. This would mean that the data series are white noise and that it makes no sense to try to identify a CEGH model. This measure of the unexplained is useful to assess whether increasing the order of the filter (number of time delays considered) improves the model or not.

4. Format of the input data.

The series of data to be analyzed must be included in a text file with the format specified below. As an example of this type of file, you can see the file "Series_BPS50.txt" that contains the series of weekly average contributions to the Bonete, Palmar and Salto Grande dams, which are the three most important in the hydroelectric system of Uruguay. This file can be downloaded from the address: https://simsee.org/downloads.html

Fig.3 shows the beginning of the aforementioned file that will serve as an example. The file is a simple text file. As a decimal separator, the point "." should be used, and a tab to separate columns (marked as gray arrows in Fig.3).

```
3	NSeries
1909	1	1	0	0	0	// año	mes	dia	hora	minuto	segundo	fecha de la primera muestra
168.5723	// Período de muestreo en horas
5216	NPuntos
52	Puntos por ciclo
	Bonete	Palmar	Salto
0	55	4	2976
1	50	3	1987
2	34	7	1297
3	42	3	799
4	24	9	589
5	17	16	506
6	16	33	865
7	13	87	1365
```

Fig. 3: Format of the data series file to analyze.

In the first line of the file, you must specify the amount of data series included. In this example they are 3.

The second line contains the temporary start time of the series, that is, the date and time of the first sample. The format is: year, month, day, hour, minute, second, all separated by tabs. The sequence "//" indicates that the rest of the line is a comment. In the example "// year month day hour minute second ..." is just a comment.

The third line specifies the sampling period in hours. In this case it says 168.5723, to indicate that the samples are one per week. As the historical data in this case are given as 52 values per year and it is a long series (100 years), the duration of the weeks in hours must be calculated so that there is no time lag. (They are not 168 hours' weeks).

The fourth line tells the number of samples in each series. In this case it says 5216. This is, there are 5216 data for each series starting the first on 1/1/1909 00:00:00.

The fifth line tells the number of points per cycle that would be advisable to use in the identification of deforming functions. In this example it says 52, and indicates that the data has a seasonality that is repeated every 52 samples.

The sixth line has a tabulator (first blank column) and then the names with which the data series are identified. In this case they are "Bonete", "Palmar" and "Salto".

From the seventh line onwards there is a value in the first column that identifies the passage of time. In the example, only the first three time steps are shown, therefore having in the first column 0, 1, 2. In the complete file of this example, the lines continue until the time step whose ordinal is 5215. In the columns corresponding to each series of data is the verified value for each time step. The ordinal (first column) can be an integer or floating-point numbers, it is only for user reference, it is not used in calculations by AnalisisSerial.

As explained above based on Fig.3, it is the simplest format supported by AnalisisSerial and is the Version 0 (zero) format.

Fig. 4 shows another example (file "series_central_36.txt" also available at https://simsee.org/downloads.html).

```
VERSION_FORMATO_SERIES:	3	// (ver. >=1 )version number. If the line does not start with "VERSION_FORMATO_SERIES" it is assumed "y = 0"
7	// N Series
2017	7	19	0	0	0		// date of first sample
1.0		// sampling time step in hours
6144		// number of samples
1		// number of points for the main cycle
1		// (ver. > 2 )number of chronicles
xxxxxxy		// (ver. > 0) serie type. x = In/Out ; y = Out
		// (ver. > 2 ) empty line
kCron:	1		// (ver. > 2) chronicle id
	Velocidad_(m/s)	Cos(Direccion)	Sin(Direccion)	Radiacion_(W/m^2)	Densidad_del_aire (kg/m3)	Temperatura_(gr C)	PotenciaPorUnidadDisponibleSinRestriccion
42935	10.52777778	0.874619707	-0.48480962	0	1.2811	4.1	-999999
42935.04167	10.36111111	0.829037573	-0.559192903	0	1.2814	3.9	-999999
42935.08333	10.05555556	0.829037573	-0.559192903	0	1.2817	3.5	-999999
42935.125	10.27777778	0.79863551	-0.601815023	0	1.2831	3.3	-999999
42935.16667	10.25	0.777145961	-0.629320391	0	1.2835	3.2	-999999
42935.20833	9.972222222	0.743144825	-0.669130606	0	1.2845	3	-999999
42935.25	9.861111111	0.75470958	-0.656059029	0	1.2867	2.5	-999999
42935.29167	9.861111111	0.766044443	-0.64278761	0	1.2878	2.2	-999999
42935.33333	10.11111111	0.777145961	-0.629320391	22	1.2876	2.3	-999999
42935.375	9.5	0.75470958	-0.656059029	148	1.2823	3.5	0.829088747
42935.41667	8.555555556	0.79863551	-0.601815023	286	1.2609	8.1	0.787180021
42935.45833	8.416666667	0.731353702	-0.68199836	439	1.2525	9.6	0.75631931
```

Fig. 4: Series format in Version 3.

As you can see, the first line begins with "VERSION_FORMATO_SERIES:" and that indicates that it is the line that defines the format. In this case, the next 3 on the same line (after a tab) is the version number. In the comments of this example it has been indicated with "(ver.> = 1)" that this line only appears in versions greater than or equal to "1" and that in the case of not finding the character string "VERSION_FORMATO_SERIES:" at the beginning of the line, it is assumed that the version is 0 (zero).

In line 8 of Fig.4 there is a string "xxxxxxy" that indicates how the data series should be treated. This classification was introduced in version 1. Series marked "x" are considered inputs (variables that explain) and can in turn be outputs (variables explained) in models with delays. The "y" series are only outputs (variables explained). In the example of Fig.4, the last series (last column) corresponds to the power generated (expressed in per unit of the installed power) and the first six series correspond to measurements of the park's weather station. It makes sense that the first 6 series (and eventually their past) explain the last series.

The lines, which in the comments were marked with "(ver.> 2)" only appear as of version 2 of the format. In version 2, the possibility of describing a set (ensemble) of chronicles (or realizations) was introduced. In Fig.4, line 7 indicates the number of members of the ensemble (in this case 1), line 9 is intentionally left blank and line 10 contains the identifier of the chronicle (or

member of the ensemble). Next to lines 9 and 10, the sample set begins, one per row as described in the example in Fig.3. At the end of the data of the first chronicle (member of the enesemble), the heading is repeated, a blank line (such as 9 in Fig.4), followed by an identification line of the chronicle (such as 10 in Fig.4) and then the samples of the new chronicle.

5. Parameters.

In Fig.1 you can see the parameters that must be set before pressing the "Create CEGH model" button.

For ease of description, the parameters are grouped into panels.

5.1. "Deformers" Panel .

Fig. 5: "Deformers" Panel.

En la Fig.5 se muestra el contenido del Panel "Deformadores". Estos parámetros determinan cómo serán calculadas las funciones no-lineales que realizan la transformación de las series a un espacio gaussiano.

- **Samples / Cycle.** This parameter is read from the data series file but can be modified in the form before generating the CEGH model.
- **Overlapping.** This parameter is integer and can be 0, 1, 2, ... N. Overlaping indicates the number of steps adjacent to each sample in which the sample should be considered. For example, if we set a value of 2, we are indicating that each sequence of samples should be considered for the moment at which each sample corresponds and for the next 2 steps and for the previous 2 steps. In some way, Overlapping relativizes the time information in which a sample occurs, reusing it in the previous and previous steps. This has a strong impact on the definition of deforming functions, given that it contributes more samples to the formation of histograms and also relativizes the occurrence of unlikely extreme events, allowing them to occur in a time environment where they occurred. in the historical data series. For example, in the series of flow data of hydraulic contributions to dams, there are some extreme events in a week of September, which if you put Overlapping = 0 will force the deforming functions to allow the occurrence of that extreme only in that specific week in which it was verified in the historical data when it is reasonable to assume that the same event could have occurred in an environment of more or less 2 weeks than the one verified in the historical series (see section 5.5).
- **Traslapping.** This parameter is an integer 0, 1, 2, ... N. Traslapping is similar to Overlapping, in that it causes the same sample to be used in more than one temporary position. Unlike overlapping, instead of being temporary positions adjacent to the original, they are positions obtained from the original by "jumps" of NPuntosPorMiniCiclo (see section 5.5).

- **Samples / MiniCycle**. It sets the amount of time steps that are used by the Traslapping mechanism to identify the moments in which the information of a sample is applied (see section .5.5).
- **N° Points.** Indicates the number of points (discretizations) that will be used to represent the deformators.
- **Format.** Specifies the precision (number of digits and decimals) that will be used to write the deformers in the output file of the CEGH model.
- **Measurement noise injection.** This panel allows to specify that, for the construction of the non-linear transformations that transform the series from the original space to a Gaussian space, it is considered that the series correspond to measurements made with a given precision, and therefore Gaussian white noise is injected with the standard deviation corresponding to the precision. The "N° of chronicles" parameter indicates the number of times each value of the original series will be considered. If "N° of chronicles = 1" only the values of the original series are considered. If N° of chronicles> 1, in addition to the original values, "N° of chronicles -1" values obtained from the originals by adding zero Gaussian white noise of zero mean value and standard deviation equal to the precision specified in the parameter "precision [pu] " will be considered.

5.2. "Filter"Panel.

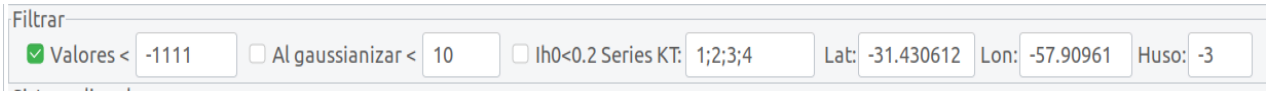

Fig. 6: "Filter" Panel

Fig.6 shows the contents of the "Filter" Panel whose content is described below.

- **Values <.** This parameter has a box that if checked, this filter is applied on the samples, and if it is not checked it is not used. If checked, the filter consists of discarding the sample sequences containing some value lower than the one specified in the box (with value -1111 in Fig.6). This serves to signal invalid data in the data series and thus disregard them. For example, in a wind speed acquisition system, when we detect that the anemometer is broken, we set -999999 for those values to be filtered. If the values were eliminated directly, the synchronism of the data regarding the time of the year would be lost.
- .**To gaussianize <.** Specifies a filter to discard samples when the value is less than the specified value when transformed into Gaussian space. This is useful so that, during the identification of the linear system (in Gaussian space), samples that correspond to extreme situations do not intervene.
- **Ih0 < 0.2 Series KT, Lat, Lon y Huso.** This filter was designed for the treatment of series of the KT clarity index (solar radiation measured on

a horizontal plane on the earth's surface / extraterrestrial solar radiation incident at the same coordinates). This allows the KT series to be analyzed in conjunction with other series (for example wind speed) and to have a selective filtration at night hours (or very low radiation), applying a filter on the KT series when the Extraterrestrial Solar Radiation $I_{h0}<0.2 kW/m^2$. If checked, a filter is applied to the series specified in the text field. In Fig.6, the filter will be applied on series 1, 2, 3 and 4 (the first four).
The geographical location is determined with the "Lat" (Latitude) and "Lon" (Longitude) fields both in decimal degrees. The "Time Zone" field must contain the time zone with which the samples are identified (from the date and time of the first sample and the sampling interval). In the example, the coordinates are those of Uruguay and the time zone indicates that the series are based on GMT-3.

5.3. "Linear System" Panel

Fig.7 shows the content of the "Linear System" panel that allows you to specify the type of linear system to be identified and its level of complexity.

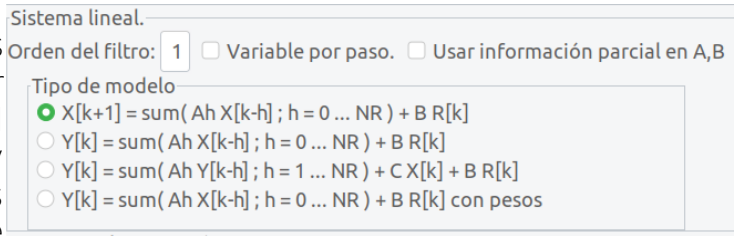

Fig. 7: "Linear system" panel.

- **Filter Order**. This parameter accepts integer values 1, 2, 3, ... N and specifies the memory (number of delay steps considered) of the recursive filter. The value 1 indicates that the filter estimates its output based on the information from the previous time step. If the value is 2, the last two steps of time are considered to project the output, if the order is 3, the last three and so on. Increasing this value implies increasing the memory of the filter and therefore the number of parameters to be estimated, and also the dimension of the state variable of the linear system. The best, most convenient value of this parameter should arise from an analysis of the results.
- **Variable per step.** If unchecked, a single linear system is identified. If checked, a linear system is identified for each step of the main cycle specified in the "Samples / Cycle" parameter in the data series file.
- **Use partial information in A, B.** This parameter allows you to specify what is done with the samples when any of the series is filtered (by any of the filtering mechanisms). If it is not marked, the entire sample is discarded (that is, the values of all the series). On the contrary, if the box is checked, it is tried to take advantage of the values of the series that are not filtered.
- **Type of model.** This subpanel allows you to select the type of model to use among those available. The first is the "classic" used for the construction of CEGHs models in SimSEE. The others are variations used

mostly for data analysis. In the latest versions of SimSEE (viie20-197) the use of option 3 was enabled. Y[k] = sum(Ah Y[k-h] ; h = 1 ... NR) + C X[k] + B R[k] . Option 2 is a particular case of 3. Option 4 is still experimental.

5.4. "*Printing options*" Panel.

Fig. 8: "Printing options" Panel.

The "Printing options" panel, see Fig.8, allows you to specify the printing to auxiliary files with the series covariances for different delay steps. Covariances are calculated on the series in Gaussian space. If the "Calculate covariance each cycle step" box is checked, they are calculated independently for each step of the main cycle (determined by the "Samples / Per cycle" parameter in the data series file). The number of maximum delays to be printed must be indicated in the "N° of steps" field. All files are generated in the results folder "{$ HOME} / SimSEE / AnalisisSerial".

5.5. *More about Overlapping y Traslapping.*

The Overlapping and Traslapping parameters allow to indicate that the effect of the same sample is assigned to the lateral steps and "jumping" NPpoints PorMinicilos steps to the right or left (future and past respectively).

If Overlapping = 3 and Traslapping = 0, the same sample will be considered in the instant of time to which it is associated, in the previous 3 and in the following 3.

If Overlapping = 3 and Traslapping = 2 with NPpoints Per Mini cycle = 10, the same sample will be considered in its time step (box k), in the previous 3 and in the next 3, and that group of 7 samples (3 + 1 + 3) will be considered centered in box k to which the sample is associated and moved to both sides of that box with displacements -20, -10, +10 and +20.

Using both parameters, the same sample will be considered in $(2\,Overlapping+1)(2\,Traslapping+1)$ boxes.

Usage examples:

- In the CEGH identification of average weekly contributions to dams: Overlapping = 3 and Traslapping = 0 can be reasonable values.
- In the identification of hourly wind speeds: Overlapping = 2 and Traslapping = 7 with NPpoints PerMinicycle = 24 are reasonable values.

The following Fig.9 tries to clarify the use of these parameters.

The first series shows the original series in which a sample has been marked with a "1" (yellow box).

The second frame shows the effect of Overlapping = 2, Traslapping = 0. The original series is expanded in 5 series, moving the original back and forth in 2 boxes.

The third table shows the effect of Overlapping = 2, Traslapping = 1, NPpoints PerMinicycle = 7. As can be seen, the group of 5 series previously expanded by Overlapping = 2 is in turn expanded into two new groups, displacing that group in 7 boxes (number of NPpoints PerMinicycle) once forward and once backwards because of having imposed Traslapping = 1.

Serie Original
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Series consideradas si Overlapping=2 y Traslapping=0	ov.	tr.
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-2	0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-1	0
0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0	0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1	0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2	0

Series consideradas si Overlapping=2 y Traslapping=1 NPuntosPorMiniciclo=7	ov.	tr.
0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-2	-1
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-1	-1
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0	-1
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1	-1
0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2	-1
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-2	0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-1	0
0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0	0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1	0
0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2	0
0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-2	1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-1	1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0	1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0	1	1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0	2	1

Fig. 9: Examples use of Overlapping and Traslapping.

6. Results File Format.

The results file is saved in the folder "{$ HOME} / SimSEE / AnalisisSerial" and has the name: "Synthesizer CEGH.txt".

This file contains the description of the deformation functions necessary to construct the non-linear functions to move from the Gaussian space to the real one and the description of the linear filter. First there are the deformation functions and then the description of the linear filter.

The following table shows an example of the beginning of the description of the deforming functions.

```
<+FUNCIONES DEFORMANTES>
NSS   3      Número de Series de Salida
NPP   52     Número de Puntos por Período
NPFD 200   Número de Puntos por Función Deformante
DurPasoSorteo    168
```

The meaning of the parameters is as follows:

- NSS = Number of Output Series. Indicates that the synthesizer in question generates an NSS series vector. In the example in the previous table, it corresponds to the output of the flow synthesizer identification for the Bonete, Palmar and Salto dams and therefore NSS is 3.
- NPP = Number of Points per Period. Deforming functions can be calculated in order to capture the seasonality of the original data in them. Depending on the type of data, the seasonality that is worth trying to capture in the same deformation functions will depend. In the example, the time series of data corresponded to the weekly average flows, to the Bonete, Palmar and Salto dams, with 52 values per year. It is clear that the histograms of the series show a marked seasonality. To reproduce the histograms with the same seasonality, the deforming functions were performed considering a different one for each week of the year and therefore in the example NPP = 52.
- NPFD = Number of points of the deforming functions. Deforming functions are described by the inverse of the cumulative probability curve. This function is described by the values it takes in a discretization carried out of the interval (0,1). NPFD is the number of points considered in that discretization.
- DurPasoSorteo = is the duration of the draw step in hours of the original data. This value is necessary to determine the behavior of the model in SimSEE. For example, if the draw step is 168 hours as in the example, and is used in a SimSEE simulation with an hourly time simulation step, the synthesizer will generate values every 168

simulation steps. In the simulation steps, SimSEE considers the interpolation of the synthesized values. It can also happen in reverse, that is to say that the simulation step of the simulation that is being carried out in SimSEE is greater than the draw step of a random source (in this case our CEGH model). SimSEE automatically manages the sub-sampling, over-sampling or synchronous sampling situations, but to do so you need to know the valid time step for each sub-model. This is the purpose of this parameter.

Continued with the description of the contents of the results file, following the heading shown in the previous table is the description of the deforming functions for each of the series. In the example, we will then have 3 blocks (one for each series) with the description of the 52 deforming functions (one for each period point) of each series. Each deforming function is described in the example by the 200 values of the inverse of the cumulative probability function.

Fig.10 shows the beginning of the description of the deforming functions of the first of the series (Bonete in the example).

serie1	Bonete							
	0.50%	1.00%	1.50%	2.00%	2.50%	3.00%	3.50%	...
paso:	1	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	2	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	3	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	4	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	5	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	7	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	8	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	...
paso:	11	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	

Fig. 10: Start of description of deforming functions.

The result file is plain text with the values separated by tabs, so if it is opened with an application like Excel, it is displayed as shown in Fig.10.

In the first row, first column is the word series followed by the ordinal (1, 2 ... NSS) to which the description corresponds. In the second column, the name that identifies the series. This name is the one that appears as "terminal name" when the model is used in SimSEE.

In the second row, the first two columns (two tabs) are free and then there is the probability to which each of the columns corresponds. There are as many columns as NPFD.

Then, for each NPP there is a row, with the word "step" in the first column and the ordinal of the point within the period to which the deforming function contained in the same row corresponds. In the following columns are the inverse values of the cumulative probability density function for each of the probabilities listed in row 2.

At the end of the block corresponding to a series, a blank row is left and the description of the deforming functions of the next series begins and so on until all the series have been described.

After the description of the deforming functions, the description of the linear filter and possible reductions of the state to be applied in SimSEE follows.

La Fig.11 shows the end of the file used as an example, with the description of the linear filter and a reduction of the state variables from 3 to 2 variables.

<+FILTRO LINEAL>										
NFRBG	3									
NSS	3									
NCOLSA	3									
Filtro A										
			S1-1	S2-1	S3-1	|	u1	u2	u3	|
serie:	1	|	7.61E-01	2.49E-02	7.64E-02	|	3.80E-01	-1.81E-01	-3.97E-01	|
serie:	2	|	1.58E-01	6.26E-01	9.50E-03	|	6.14E-01	2.53E-01	1.74E-01	|
serie:	3	|	1.21E-01	-3.31E-02	7.80E-01	|	1.82E-01	-4.77E-01	2.43E-01	|
nVE	2									
nd1	5	Estado_H1	0.5	0	0.5	Estadolnicial	0			
probs	0.2	0.2	0.2	0.2	0.2					
nd2	5	Estado_H2	0	1	0	Estadolnicial	0			
probs	0.2	0.2	0.2	0.2	0.2					

Fig. 11: Description of the linear filter and the state reduction.

At the beginning in Fig.11, the text "<LINEAR FILTER>" is in the first column, indicating that the description of the linear filter begins.
The next three rows define the values of NFRBG, NSS and NCOLSA, which in the example are 3, 3 and 3 and whose meanings are:

- NFRBG = number of sources of Gaussian white noise. At present, by construction, this number coincides with the number of series, but it could be different if the identification algorithm is changed later.
- NSS = Number of series (model outputs)
- NCOLSA = Number of columns of the main matrix of the linear filter. This number is equal to NSS multiplied by the number of delays in the time taken into account by the linear filter. In the example NCOLSA = 3 = NSS means that the filter in question only considers 1 (one) delay of the signals over time.

After this heading, there is a blank line and another that begins with the word "Filter A" announcing that the coefficients of the linear filter come.

The linear filter is identified by two matrices A and B, each of NSS rows.

Matrix A has NCOLSA columns and matrix B has NFRBG columns. In the table, matrix A is headed by the texts: "S1-1", "S2-1" and "S3-1" that indicate that their coefficients multiply the value of the series 1, 2 and 3 in the previous step "-one".

In the example, matrix B is headed by "u1", "u2" and "u3" and multiply the values of the Gaussian white noise sources.

Each row of the matrices is preceded by a row identifier of the type "series: 1 |" (first three columns of the table) followed by the coefficients.

Each row describes how to calculate the next filter output for each series. As an example, using the coefficients of row k we can write:

$$s_k(t) = \sum_{i=1}^{i=NSS} \sum_{j=1}^{j=NRT} A[k,(i-1)*NRT+j] * s_k(t-jT) + \sum_{h=1}^{h=NFRBG} B[k,h] * u_h(t)$$

Where:

- $s_k(t)$, is the filter output for the series ***k.***
- ***NRT*** is the amount of time delays considered and can be calculated as: ***NRT = NCOLSA / NSS***. (is 1 in the example).
- $u_h(t)$, are the Gaussian white noise inputs of variance 1.

Continuing with the description of the results file, after the rows corresponding to the matrices of the filter, a blank line comes and then begins the description of the state variables to be used if the CEGH model is used for a Stochastic Dynamic Optimization. By default, AnalisisSerial writes the description of the state variables as a variable for each column of matrix A, indicating a discretization of 5 for each variable and assigning a 20% probability to each discretization. In that case, the state of the linear filter has the same dimension in Optimization as in Simulation.

In practice, generally this part of the file must be edited manually to reflect the reduction of the state space that is to be carried out in the Optimization stage in SimSEE, as a way to combat the Bellman's Curse of Dimensionality. For example, in Fig.11 it is shown that instead of 3 state variables, 2 will be used for optimization, that the first of these variables is discretized at 5 points in the state space, it is called "State_H1" and it is defined as: State_H1 = 0.5 * Bonete + 0.0 * Palmar + 0.5 * Salto (always operating on the signals in Gaussian space).

The variable State_H2, will also be discretized in 5 points and is defined as: State_H1 = 0.0 * Bonete + 1.0 * Palmar + 0.0 * Salto.

By default, AnalisisSerial writes a description of the state variables for optimization that matches the state of the linear filter imposing 5 (five) discretization points for each variable. For example, in the model of the example in Fig.11, at the exit of the AnalisisSerial program, it would say nVE = 3 and then comes the description of the three variables with the same name as the series and with 5 discretization points each with probabilities 0.2 for each point and the coefficients of the reducing matrix would be such that it is identity.

The state variables have by construction normal distribution N (0,1). The division of the range of each variable into "points" for the discretization of the state space implies that each point is the "center" of a probability band. The value indicated in the model is the area (cumulative probability) associated with the probability band, and the point of discretization is considered in the value of the variable that leaves the same area within the band on each side of the point.

To fix ideas, Fig.12 shows the curve corresponding to the cumulative probability density (CDF) of a variable ***x*** with Normal distribution of mean zero and unit variance. The dotted vertical lines (green) separate the 0.2 accumulated probability bands. The circles with the vertical arrows identify

the central points in each band that correspond to the values considered as discretization points of the variable when used in the SimSEE Stochastic Dynamic Programming algorithm.

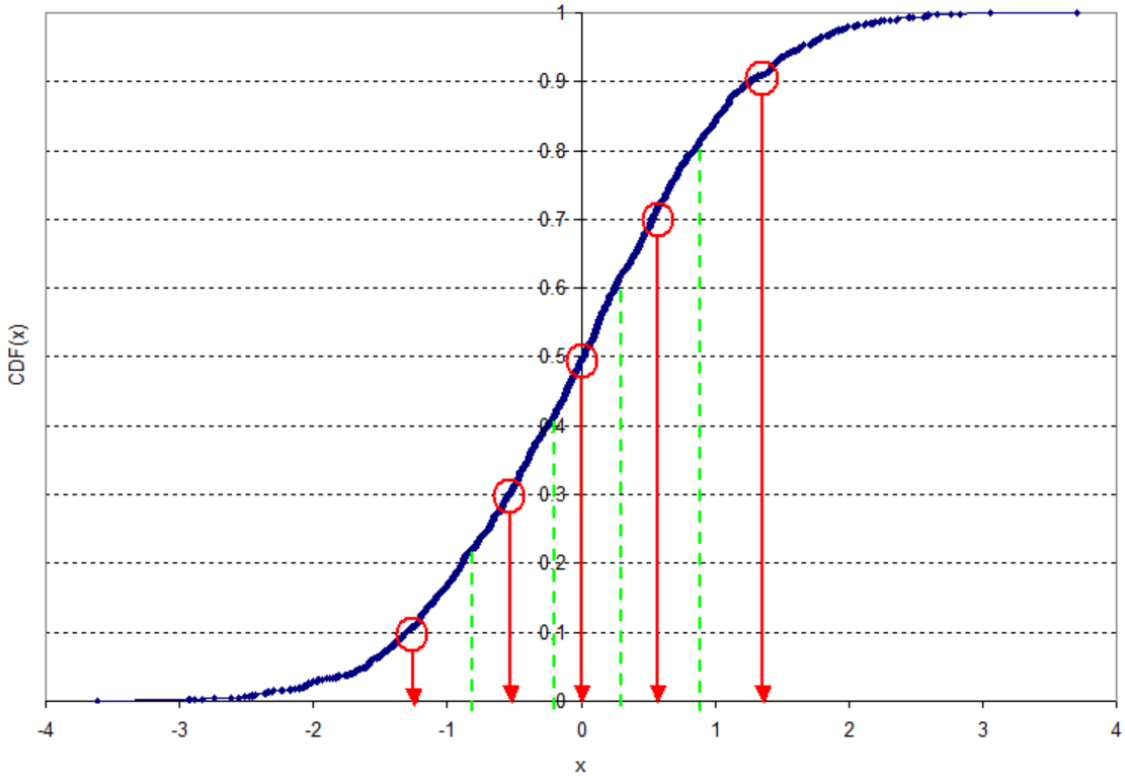

Fig. 12: Example of partitioning into 5 equiprobable bands.

7. Printing of covariance files.

Pressing the "Create CEGH" button creates additional files in the output folder, which can be useful for analyzing the modeled signals.

If $\{X_j\}$ is the series of gaussianized signals, the delay covariance matrix k sis defined as: $\Sigma_{XX}(k)=\langle X_j X^T_{j-k}\rangle$. All files are plain text, separated by tabs. The file "covars_k_gaussianas.xlt" contains the matrix coefficients $\Sigma_{XX}(k)$ for the values of k from k = 0 to "No. of steps -1". If the "Calculate covariance every cycle step" box is checked, additional files are generated, with names of the type "SerieA (x) SerieB.xlt" with the coefficients of the covariance between "SerieA" and "SerieB", for delays from 0 to "N° of steps - 1" and for each of the steps in the cycle determined by "Samples / Per cycle".

8. Technical notes.

8.1. Matrix B calculation of the linear system.

For the calculation of the matrix B of the filter, two options are originally allowed. A decomposition by the Grand Smith method of the noise vectors or the Cholesky decomposition of the co-variance matrix of the noise vectors. Recently the option "Algebraica + Cholesky" was added, which is the one that is active and the possibility of users choosing the calculation method of matrix B was disabled. The current method calculates the matrix $\langle BB^T \rangle$ with algebraic operations from the covariance matrix of the data series and the matrix A identified for the linear filter by least squares, and then B is calculated by Cholesky decomposition. This form of calculation guarantees that the variances of the linear filter outputs are unitary. The above methods use the residuals of the identification of A for the estimation of $\langle BB^T \rangle$ what does not ensure that the variance was unitary since said residues are not 100% uncorrelated.

The three forms of calculation achieve a matrix B that breaks down noise vectors into statistically independent vectors, but the matrices are not the same, as there are many solutions to the problem. It was decided to leave the option of changing the method for the common user disabled (Advanced users can re-enable the option in the project sources of the application and recompile it.).

Bibliografía

1: Ruben Chaer, Fundamentos de modelo CEGH de procesos estocásticos multivariables., 2011