



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
MÁSTER EN INGENIERÍA INDUSTRIAL

BASIC ANALYTIC SYSTEM FOR INTERVAL VALUED DATA

Autor: Pablo Maceda Dal-Re
Director: Carlos Maté Jiménez

Madrid
Julio, 2018

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESIS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. Pablo Maceda Dal-Re

DECLARA ser el titular de los derechos de propiedad intelectual de la obra:

Basic Analytic System For Interval-Valued Data, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

- a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar “marcas de agua” o cualquier otro sistema de seguridad o de protección.
- b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
- c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
- d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
- e) Asignar por defecto a estos trabajos una licencia Creative Commons.
- f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente*).

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

- a) Que la Universidad identifique claramente su nombre como autor de la misma
- b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
- c) Solicitar la retirada de la obra del repositorio por causa justificada.
- d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

- a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
- b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
- c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

- d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusiva del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 17. De Julio de 2018.

ACEPTA



Fdo.....

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título
Basic Analytic System For Interval-Valued Data
en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el
curso académico 2017-2018 es de mi autoría, original e inédito y
no ha sido presentado con anterioridad a otros efectos. El Proyecto no es
plagio de otro, ni total ni parcialmente y la información que ha sido tomada
de otros documentos está debidamente referenciada.



Fdo.: Pablo Maceda Dal-Re Fecha: 17/ 07/ 2018

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO



Fdo.: Carlos Maté Jiménez Fecha: 16, 07, 2018



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
MÁSTER EN INGENIERÍA INDUSTRIAL

BASIC ANALYTIC SYSTEM FOR INTERVAL VALUED DATA

Autor: Pablo Maceda Dal-Re
Director: Carlos Maté Jiménez

Madrid
Julio, 2018

BASIC ANALYTIC SYSTEM FOR INTERVAL-VALUED DATA

Autor: Maceda Da-Re, Pablo.

Director: Maté Jiménez, Carlos.

Entidad Colaboradora: ICAI- Universidad Pontificia Comillas.

RESUMEN DEL PROYECTO

INTRODUCCIÓN

El análisis de datos está a la orden del día del Business Intelligence y es esencial en las nuevas tendencias, como Big Data, Internet de las cosas (IoT), la nube o ciudades inteligentes. La toma de decisiones en el mundo real incorpora diferentes tipos de datos.

Hasta ahora, los datos simples o numéricos únicos han sido el paradigma de la estadística, el Machine Learning y la Inteligencia Artificial, las numerosas herramientas y softwares existentes cubren muchos métodos consensuados para trabajar y analizar datos simples. Sin embargo, están surgiendo nuevos tipos de datos. Los datos de intervalo son los compuestos por un intervalo típico con límites inferiores y superiores, como los máximos y mínimos semanales en finanzas (mercados FOREX, acciones o materias primas ...), temperaturas mínimas y máximas diarias, valores de tensión cardíaca y muchas otras magnitudes que no proporcionan una visión completa con un único valor simple como estamos acostumbrados.

Los datos de intervalo han existido durante mucho tiempo, de hecho, la mayoría de los valores que obtenemos en el mundo real con instrumentos de medición y cálculos son intervalos. Los valores simples que normalmente se usan incluyen algún valor de error del instrumento, algunos cálculos que convierten un valor específico en un intervalo con un límite superior e inferior. Además, hay muchos datos que no proporcionan la información completa con un solo número. Los datos de intervalo, aunque de ninguna manera son perfectos, brindan una información más amplia que los datos simples para los elementos de información que no se pueden explicar con un solo número.

Como se explica en Mate (2012) [1], el primer análisis de intervalo comenzó en la década de 1960 con Ramon Moore, pero el enfoque inicial se centró principalmente en los resultados computacionales surgidos de los datos de intervalo. A partir de entonces, los intervalos han ganado mucho interés en el mundo de la investigación, especialmente en el siglo XXI, donde han llegado la mayoría de sus aplicaciones y análisis.

Aunque esto puede parecer un tema trivial ya que los datos de intervalo nos han acompañado durante mucho tiempo, su análisis y estudio todavía está en desarrollo y muchos avances solo se han producido a principios del siglo XXI o incluso en nuestra década actual, dejando aún muchos huecos sin rellenar y muchos métodos similares para estudiar y trabajar con ellos.

La visualización es un campo abierto que aún debe ser consensuado con respecto a los datos de intervalo. Si bien los datos únicos tienen gráficos y figuras claramente estandarizados que se usan comúnmente, los intervalos son mucho más difíciles de representar y analizar visualmente, especialmente en los gráficos bidimensionales a los que estamos acostumbrados.

Se han hecho algunos intentos para representar intervalos como, por ejemplo, series temporales de intervalo o un tipo de diagrama de dispersión para dos variables de datos con valores de intervalo diferentes. Este proyecto incluirá algunas de estas representaciones existentes y creará nuevas representaciones de datos de intervalo para un mayor análisis visual y comprensión.

La regresión es un área donde la investigación ha avanzado recientemente. Los intervalos, a diferencia de los datos únicos, tienen muchos valores significativos (centro, radio, límite inferior y límite superior) lo que hace que los métodos clásicos de regresión sean complicados de aplicar. Los primeros métodos básicos creados para la regresión de intervalos fueron de Billard y Diday en 2000 y 2002[2]. Más tarde fueron recogidos y mejorados por Neto y Carvalho en 2008 y 2010[3].

Muchos de estos métodos, especialmente los más antiguos, usan análisis de regresión de valor único aplicados a diferentes parámetros del intervalo; sin embargo, los más recientes, como Sinova et al en 2012 [4] o Souza et al en un artículo en 2017[5], intentan profundizar y utilizar técnicas como la aritmética de los intervalos o su parametrización.

Para este proyecto, se probará e incluirá una amplia gama de métodos de regresión en la herramienta, algunos de los originales de Billard y Diday y Neto y Carvalho, y algunos métodos más recientes y avanzados que se han propuesto en los últimos dos años.

Las series temporales de intervalo son probablemente el área más investigada de los datos de intervalo. Las aplicaciones en los campos de negocios y finanzas, donde los datos significativos, como los valores de la bolsa o la fluctuación de la moneda, se pueden representar como series temporales de intervalo, es lo que más interés ha producido.

Para las series temporales de intervalo, a pesar de que ha sido el área más investigada de los datos de intervalo de valores, la mayor parte de su análisis se ha centrado en los métodos de previsión. La previsión es el análisis más útil para las aplicaciones financieras típicas de las series de temporales de intervalo.

Este proyecto no se centrará en la previsión, ya que es un tema muy amplio y cubierto, y más en la representación de la serie temporal de intervalo y otros análisis diferentes.

MATLAB ha implementado muy poco en el análisis de intervalo. En 1999 se implementó una librería (INTLAB) con aritmética de intervalos, pero eso ha sido todo lo que MATLAB ha avanzado con respecto a intervalos. A diferencia de R, donde se han creado otras librerías con funciones de regresión o visualización, MATLAB todavía tiene un largo camino por recorrer con respecto al Análisis de Intervalos, dando lugar a que este proyecto se desarrolle con esta necesidad.

Este proyecto desarrollará un sistema nuevo y original para analizar datos de intervalo que combine y compare los diferentes métodos existentes y los converjan en una sola herramienta. Esta herramienta debería ayudar al usuario a visualizar y analizar una nube, previamente incomprensible, de puntos de datos de intervalo que tengan en cuenta los métodos y alternativas

ya existentes, incluidas nuevas formas creativas de mostrar los resultados y facilitar la tarea de los no expertos usuarios de programación para trabajar y entenderlos.

METODOLOGÍA

Este proyecto tiene dos partes principales con diferentes metodologías:

1- Lectura e investigación de artículos existentes, librerías de codificación y tesis previas que puedan proporcionar herramientas útiles para el análisis, la regresión, la visualización o la medición de datos de intervalos. Algunos serán recomendados y proporcionados por el director, otros encontrados a través de investigaciones propias.

2- Programación y uso de MATLAB. Esta parte es lo que llevará la mayor parte del tiempo: primero la programación de las funciones necesarias para las diferentes funcionalidades de la herramienta; una vez que esas funciones se completen, la creación de la herramienta que, debido a sus singularidades, forzó la modificación y el ajuste de algunos detalles de las funciones que previamente se habían creado para que coincidan con los requisitos del diseñador de la aplicación. Para esta parte, la mayor fuente de información ha sido la documentación y la sección de “answers” de la propia página web de MATLAB, Mathworks [6].

RESULTADOS

El resultado final de este proyecto es la herramienta Basic Analytic System for Interval Valued Data (BASIVD) que se muestra en la Figura 1. Esta herramienta le dará al usuario la posibilidad de realizar diferentes análisis y representar diferentes figuras tanto para series temporales como para datos de corte transversal.

Esta herramienta es fácil de utilizar y permite ser usada por gente que no esté acostumbrada a la programación o a MATLAB. Su instalación es sencilla y puede realizarse como una App de MATABL o instalada como un programa más en tu ordenador.

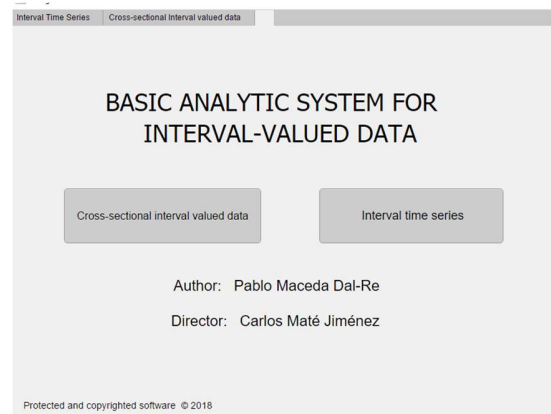


Figura 1 Portada de BASIVD

La herramienta tiene varias funcionalidades, las dos principales son la visualización y el análisis. Respecto a la visualización, se pueden representar diferentes gráficas tanto para series temporales como para datos truncados dependiendo del número de variables y de que busca ver el usuario. Un ejemplo de esto es lo mostrado por la Figura 2, una gráfica sacada por la herramienta en la que se muestran todos los intervalos de una variable y sus intervalos principales en la parte derecha.

En la parte de análisis, el principal realizado por esta herramienta es la regresión. Incluye tipos diferentes de modelos de regresión con 10 medidas de bondad de ajuste que pueden ser seleccionadas por el usuario. La herramienta retorna los valores de regresión de cada modelo para

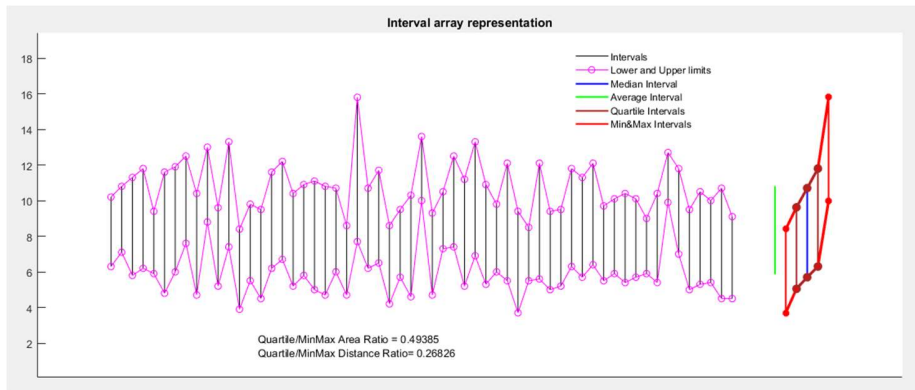


Figura 2 Ejemplo de visualización de BASIVD

cada medida seleccionada y si se trata de una regresión simple se muestran las líneas de ajuste del modelo como las que se aprecian en la Figura 3.

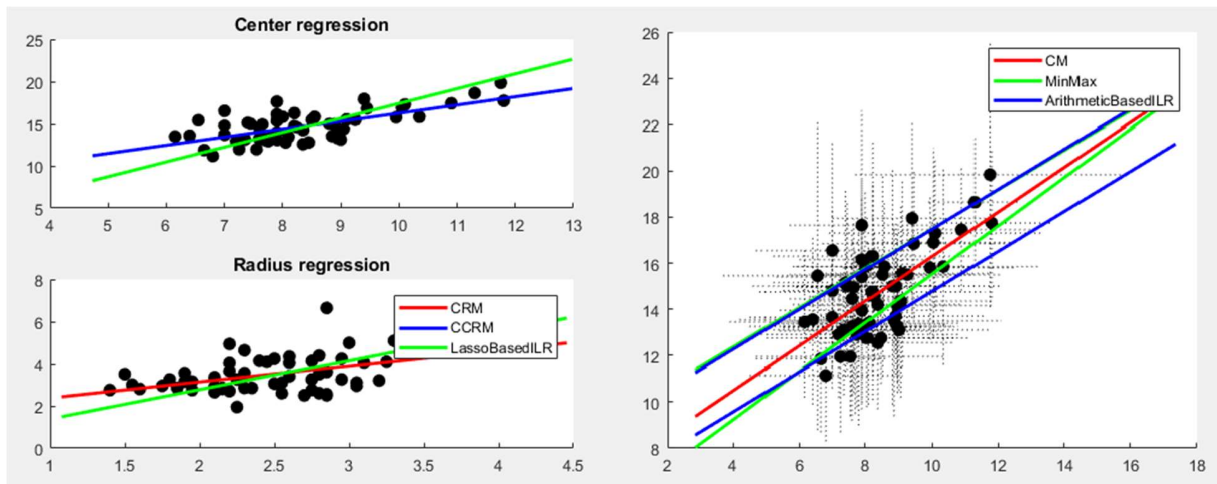


Figura 3 Ejemplo de regresión de BASIVD

CONCLUSIONES

El resultado final de este proyecto es la herramienta BASIVD que incluye dos principales componentes: visualización y regresión. En la parte de visualización se han incluido gráficas y representaciones originales e innovadoras para intentar solucionar la complejidad de representar datos por intervalos. Para la parte de regresión, una de las más completas en la herramienta, se realizaron análisis para comparar los diferentes métodos. Las conclusiones sacadas de este análisis es que no existe un método claramente superior al resto. Los modelos más modernos y avanzados realizan una predicción ligeramente mejor que los demás, pero esto lo realizan a cambio de complejidad y poder computacional. Por tanto, depende de las necesidades del usuario si prefiere un modelo ligeramente más preciso y complejo o le sirve una proyección buena con un modelo más sencillo.

REFERENCIAS

- [1] Carlos Maté Jiménez, El análisis de intervalos: aplicación a la ingeniería, June 2012, Anales.
- [2] Billard, L. and Diday, E, Regression analysis for interval-valued data., 2000, Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies.
- [3] Eufrásio de A. Lima Neto and Francisco de A.T. de Carvalho, Centre and Range method for fitting a linear regression model to symbolic interval data, 2008, Computational Statistics and Data Analysis.
- [4] Beatriz Sinova and Ana Colubi and María Ángeles Gil and Gil González-Rodríguez, Interval arithmetic-based simple Basic analytic system for interval-valued linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric, 2012, Information Sciences.
- [5] Leandro C. Souza and Renata M.C.R. Souza and Getúlio J.A. Amaral and Telmo M. Silva Filho, A Parametrized Approach for Linear Regression of Interval Data, 2017, Knowledge-Based Systems.
- [6] <https://www.mathworks.com>

BASIC ANALYTIC SYSTEM FOR INTERVAL-VALUED DATA

Author: Maceda Da-Re, Pablo.

Director: Maté Jiménez, Carlos.

Collaborating Institute: ICAI- Comillas Pontifical University.

SUMMARY OF THE PROJECT

INTRODUCTION

Analytics is at the core of Business Intelligence and is essential in new trends such as Big Data, Internet of Things, Cloud Computing or Smart Cities. Decision-making in the real world incorporates different kinds of data.

Single or crisp data have been the paradigm until now in Statistics, Machine Learning and Artificial Intelligence, the numerous existing tools and software cover many agreed upon methods for working and analysing crisp data. However, new kind of data are arising. Interval data is a typical interval with lower and upper bounds such as weekly highs and lows in finance (FOREX, stock or commodity markets...), daily minimum and maximum temperatures, heart tension values and many other magnitudes that do not give the full picture with only a crisp value data point like we are used to.

Interval values have existed for a very long time, in fact most of the values that we obtain in the real world with measure instruments and calculations are, in fact, intervals. The crisp values that are normally used include some instrument error value, some calculations that turns a specific value into an interval with an upper and lower limit. On top of that, there are plenty of data that does not give the full information with only a single number. Interval valued data, while being by no means perfect, gives a wider information than crisp data for information items that cannot be explained with a single number.

As explained in Mate (2012) [1] the first interval analysis started in the 1960s with Ramon Moore, but the initial focus was mainly in computational results emerged from interval values. From then on, intervals have gained a lot of traction in the research world specially in the 21st century where most of its applications and analysis have arrived.

Even though this may seem a trivial issue as interval data has been around us for long time, its analysis and study is still in development and many breakthroughs have occurred only in the early twenty-first century or even in our current decade, leaving still many stones unturned and many similar methods to study and work with.

Visualisation is an open field yet to be agreed upon regarding interval data. While crisp data has clear standardised charts and figures that are commonly used, intervals are way harder to represent and to analyse visually, especially in the two-dimensional plots that we are used to.

Some attempts have been made to represent intervals for Interval time series or or a type of scatter plot for two different interval valued data variables. This project will include some of these existing

plots and create some new interval valued data representations for further visual analysis and understanding.

Regression is one area where research has advanced recently. Intervals, unlike crisp data, have many significant values (centre, radius, lower limit and upper limit) making classical regression methods complicated to apply. The first ever basic methods created for interval regression were by Billard and Diday in 2000 [2] and in 2002. Later they were picked up and improved by Neto and Carvalho in 2008 [3] and 2010.

Many of these methods, specially the earliest ones, use crisp-value regression analysis applied to different parameters of the interval; the most recent ones however, like Sinova et al in 2012 [4] or Souza et al in a paper in 2017 [5] try to go deeper and use techniques like the arithmetic of intervals or its parametrization.

For this project, a wide range of regression methods will be tested and included in the tool, some of the original ones from Billard and Diday and Neto and Carvalho and some more recent and advanced methods that have been proposed in the last couple of years.

Interval time series are probably the most researched area of interval valued data. Applications in the business and finance fields, where significant data like stock exchange values or currency fluctuation can be represented as interval time series, are where most of the interest has been drawn from.

For interval time series, even though it has been the most researched area of interval valued data, most of its analysis has focused on methods for forecasting. Forecasting is the most useful analysis for the typical finance applications of Interval time series.

This project will not focus on forecasting, as it is a very broad and covered topic, and more on Interval time series representation and other analysis to be included in the tool.

MATLAB has very little implemented on interval value data analysis. In 1999 a toolbox (INTLAB) was implemented with interval arithmetic but that is as long as MATLAB has gone so far regarding interval valued data analysis and computation. Unlike R, where other libraries have been created with regression or visualization functions exist, MATLAB still has a long way to go regarding Interval Analysis giving room for this project to work with this need.

This project will develop a new and original system to analyse interval-valued data combining and comparing the different existing methods and converging them in a single tool. This tool should help the user in the visualisation and analysis of a previously incomprehensible cloud of interval-valued data points considering the already existing methods and alternatives, including new creative ways of showing results and making it easy for the not-well-versed-in-programming user to work with and understand.

METHODOLOGY

This project has two main parts with different methodologies:

- 1- Reading and researching of existing papers, articles, coding libraries and previous thesis which could provide useful tools for interval valued data analysis, regression, visualisation or measurement. Some will be recommended and provided by the director, others found through own research.
- 2- Coding and usage of MATLAB. This part is what would take most of the time: first the programming of the functions needed for the tool's different functionalities; once those functions are completed the creation of the tool itself which, because of its singularities, forced the modification and tuning of some details of the previously coded functions to match the requirements of the App designer. For this part, the main source of information has been the documentation and answer page of MATLAB's own web page Mathworks [6].

RESULTS

The final result for this project is the Basic Analytic System for Interval Valued Data (BASIVD) tool shown in Figure 1. This tool will give the user the possibility to run several analysis and ask for different plots for ITS and Cross-sectional interval valued data.

This tool is user friendly and allows people that are not used to programming or MATLAB to be able to use it. Its installation is simple and can be used as a MATLAB App or as a standalone application installed in your computer.

The tool has many functionalities, the two main ones are visualization and analysis. On the visualization

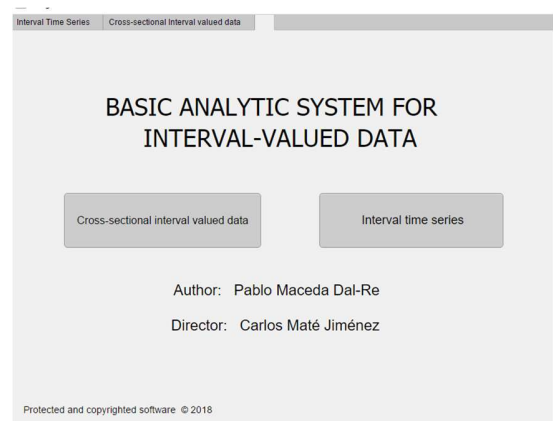


Figure 1 BASIVD main page.

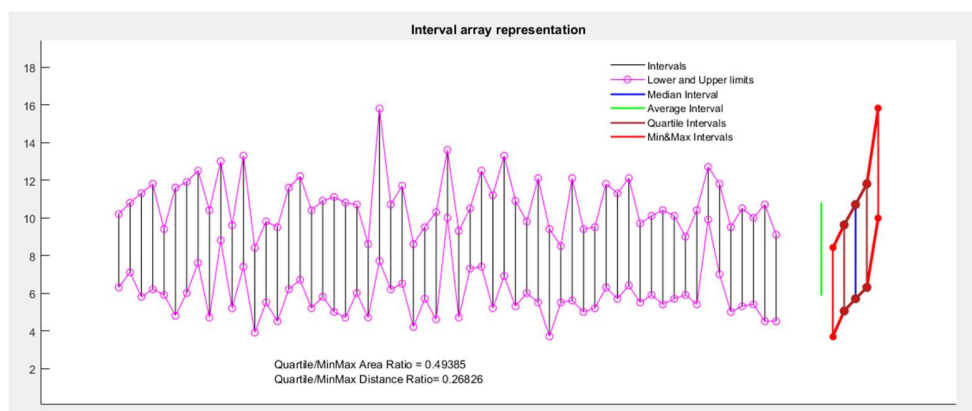


Figure 2 Visualisation example from BASIVD

part, many different plots can be shown for either ITS or Cross-sectional data depending on the number of variables and what the user wants to see.

Figure 2 shows an example of a plot produced by the tool of an interval variable with its main intervals and some dispersion variables shown.

On the analysis part, the main one produced by the tool is regression. It includes 8 different types of regression models and 10 different goodness of fit variables to be chosen by the user for the analysis. The tool gives the values of the regression models for every test selected and if it is a simple regression analysis it returns the fit lines for every method as seen in Figure 3.

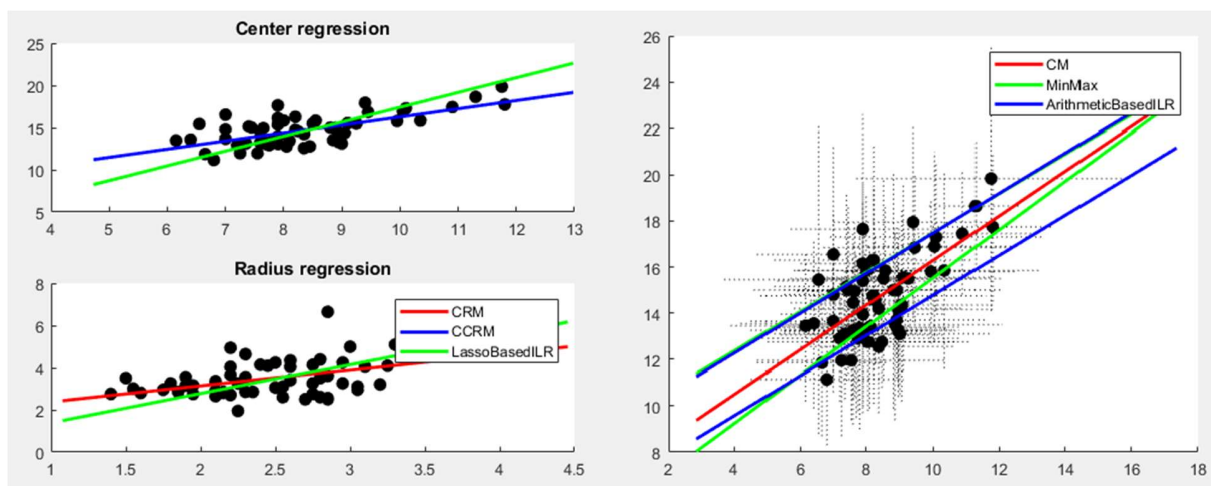


Figure 3 Regression example output by BASIVD

CONCLUSIONS

This project final product is the BASIVD tool which includes two main fields of work: visualisation and regression. For the visualisation part some innovative and original plots have been created to deal with the complexity of representing interval-valued data. Regarding regression, one of the most complete parts of the tool, some analysis was made of the different methods included on it. The conclusions drawn on this part was that there is not a clear better model. The most modern and advanced ones produced slightly better but they did that on more complexity and computational power. In the end, it is up to the user if they desire a slightly better performance with more complexity or if does prefer a simpler method with a good enough approximation.

REFERENCES

- [1] Carlos Maté Jiménez, El análisis de intervalos: aplicación a la ingeniería, June 2012, Anales.

- [2] Billard, L. and Diday, E, Regression analysis for interval-valued data., 2000, Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies.
- [3] Eufrásio de A. Lima Neto and Francisco de A.T. de Carvalho, Centre and Range method for fitting a linear regression model to symbolic interval data, 2008, Computational Statistics and Data Analysis.
- [4] Beatriz Sinova and Ana Colubi and María Ángeles Gil and Gil González-Rodríguez, Interval arithmetic-based simple Basic analytic system for interval-valued linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric, 2012, Information Sciences.
- [5] Leandro C. Souza and Renata M.C.R. Souza and Getúlio J.A. Amaral and Telmo M. Silva Filho, A Parametrized Approach for Linear Regression of Interval Data, 2017, Knowledge-Based Systems.
- [6] <https://www.mathworks.com>

*A mis padres,
como broche final de esta etapa de mi vida,
por vuestro esfuerzo, amor y sacrificio para
darme las mejores oportunidades posibles.
Siempre os estaré agradecido.*

Agradecimientos

Agradezco primero a Carlos, por darme la oportunidad de realizar este proyecto y por su imprescindible ayuda y paciencia durante este largo año. Ha sabido ir poniéndome retos, haciéndome ir un paso más allá, e informándome de las nuevas publicaciones para poder realizar un trabajo completo y actualizado.

Agradezco también a mis padres por haberme dado una educación tanto académica como moral inmejorable y haberme apoyado en todo durante estos años. También al resto de familia y amigos que me han acompañado a lo largo estos años.

Por último, agradezco a todos los profesores que he tenido durante la carrera por su pasión por enseñar y su paciencia con los alumnos en este proceso de convertirnos en ingenieros.

DOCUMENTO I



MEMORIA



Table of Contents

1. Introduction	9
1.1. State of The Art	10
1.1.1. Visualisation	11
1.1.2. Regression	12
1.1.3. Interval time series (ITS)	13
1.1.4. Intervals in MATLAB	14
1.2. Motivation	14
1.3. Objectives	15
1.4. Work Methodology	16
1.5. Resources	16
2. Visualisation	21
2.1. Cross-sectional data visualisation	21
2.1.1. Single variable	22
2.1.2. Multiple variables	26
2.2. ITS Visualization	29
3. Regression Analysis	33
3.1. Interval Linear Regression models	34
3.1.1. Centre Method (CM)	34
3.1.2. MinMax method	36
3.1.3. Centre and Range Method (CRM)	38
3.1.4. Constrained Centre and Range Method (CCRM)	40

3.1.5. Arithmetic-based simple linear regression (ABSLR)	42
3.1.6. Linear regression based on Lasso technique	45
3.1.7. Constrained centre and range joint method (CCRJM)	48
3.1.8. Parametrized Approach for Linear Regression of Interval Data	50
3.2. Goodness of fit test	52
3.2.1. iR^2	52
3.2.2. RMSE	53
3.2.3. MAPE	53
3.2.4. $iARV$	54
3.2.5. $iUTheil$	54
3.2.6. Coverage Rate	55
3.2.7. Efficiency Rate	55
3.2.8. Hausdorff Distance	56
3.2.9. Nucleus Distance	56
3.2.10. Bertoluzza Distance	57
3.3. Testing and comparison of different methods	58
3.3.1. Simple regression examples	58
3.3.2. Multiple regression example	62
4. Basic Analytic System for Interval-Valued Data	65
4.1. Cross-sectional interval valued data tab	66
4.1.1. CSIVD Visualisation tab	67
4.1.2. CSIVD Analysis tab	68
4.1.3. CSIVD Sorting tools tab	69
4.2. Interval time series tab	71
4.2.1. ITS Visualisation tab	72
4.2.2. ITS Analysis tab	73
4.2.3. ITS Cut tool tab	75

4.3. Tips and installation guide for the tool	76
4.3.1. Tips	76
4.3.2. Installation and running	77
4.3.2.1. MATLAB owners	77
4.3.2.2. Non-MATLAB owners	78
5. Conclusions	81
5.1. Visualisation	81
5.2. Regression	81
Bibliography	83

List of Figures

1.1.	Interval value data visualisation example	11
1.2.	Interval value regression model example	12
1.3.	Interval time series forecasting example	13
1.4.	Interval time series visualisation example	14
1.5.	MATLAB App Designer design view	17
1.6.	MATLAB App Designer code view	17
1.7.	MATLAB App Compiler	18
2.1.	Graphical representation of all intervals on a single variable	22
2.2.	Main intervals plot	23
2.3.	All intervals plus main intervals plot	24
2.4.	Intervals ordered by center value	25
2.5.	Three dimensional histogram of centers and radii	26
2.6.	Interval scatter plot for two variables using crosses	27
2.7.	Interval scatter plot for two variables using rectangles	28
2.8.	Classic crisp data matrix plot	28
2.9.	Interval matrix plot	29
2.10.	Multiple ITS of high and low temperatures in different cities during two years	30
2.11.	Variation of interval parameters from t_{k-1} to t_k	30
2.12.	Average and median intervals for multiple ITS	31
3.1.	Centre Method regression model example	36
3.2.	MinMax Method regression model example	38

3.3.	CRM method regression model example	40
3.4.	CCRM and CRM methods regression model comparison example	42
3.5.	ABSLR method regression model example	46
3.6.	Lasso and CCRM methods regression model comparison example	48
3.7.	Illustration of the coverage and efficiency rates	55
3.8.	Graphical representation of all model fits	58
3.9.	Median and average interval from Y and \hat{Y} of all the models	59
3.10.	Intervals for second simple regression example	60
3.11.	Prediction intervals for multiple linear regression example	62
4.1.	Home screen for BATIVD	65
4.2.	Visualisation screen for CSIVD	66
4.3.	Successful file upload	67
4.4.	Analysis screen for CSIVD	68
4.5.	Output regression methods table	69
4.6.	Sorting tools screen for CSIVD	70
4.7.	Sort by categorical variables return index	71
4.8.	Success message from Sort Dataset by threshold value . .	72
4.9.	Visualisation screen for ITS	72
4.10.	Analysis screen for ITS	74
4.11.	ACF and PACF of centres and radii of an ITS	75
4.12.	Cut tool screen for ITS	76
4.13.	MATLAB app installer	78
4.14.	MATLAB app installer	78
4.15.	Installation as a standalone application	79

Chapter 1

Introduction

ANALYTICS is at the core of Business Intelligence and is essential in new trends such as Big Data, Internet of Things, Cloud Computing or Smart Cities. Decision-making in the real world incorporates different kinds of data.

Single or crisp data have been the paradigm until now in Statistics, Machine Learning and Artificial Intelligence, the numerous existing tools and software cover many consensuated methods for working and analysing crisp data. However, new kind of data are arising. Interval data is a typical interval with lower and upper bounds such as weekly highs and lows in finance (FOREX, stock or commodity markets...), daily minimum and maximum temperatures, heart tension values and many other magnitudes that do not give the full picture with only a crisp value data point like we are used to.

Even though this may seem a trivial issue as interval data has been around us for long time, its analysis and study is still in development and many breakthroughs have occurred only in the early twenty-first century or even in our current decade, leaving still many stones unturned and many similar methods to study and work with.

This project will develop a new and original system to analyse interval-valued data combining and comparing the different existing methods and converging them in a single tool. This tool should help the user in the visualisation and analysis of a previously incomprehensible cloud of interval-valued data points taking into account some of the already existing methods and alternatives, including new creative ways of showing results and making it easy for the not-well-versed-in-programming user to work with and understand.

1.1 State of The Art

Interval values have existed for a very long time, most of the values that we obtain in the real world with measure instruments and calculations are, in fact, intervals. The crisp values that are normally used include some instrument error value, some calculations that turns a specific value into an interval with an upper and lower limit. On top of that, there are plenty of data that does not give the full information with only a single number; if we think of a temperature on a day in a certain place, we can maybe talk about the average, however, if one place is a dessert and the other is a tropical jungle, they may have the same average temperature but the difference from maximum and minimum can be very different. Same happens if you analyse only the maximum or minimum value of the day. Whichever crisp value chosen in this example, some information is being lost. Interval valued data, while being by no means perfect, gives a wider information than crisp data for information items that can not be explained with a single number.

As explained in Mate(2012) [1] the first interval analysis started in the 1960s with Ramon Moore, but the initial focus was mainly in computational results emerged from interval values. From then on,

intervals have gained a lot of traction in the research world especially in the 21st century where most of its applications and analysis have arrived. In the following sections, some background will be given on different areas of interval value data analysis.

1.1.1 Visualisation

Visualisation is an open field yet to be agreed upon regarding interval data. While crisp data has clear standardised charts and figures that are commonly used, intervals are way harder to represent and to analyse visually, specially in the two-dimensional plots that we are used to.

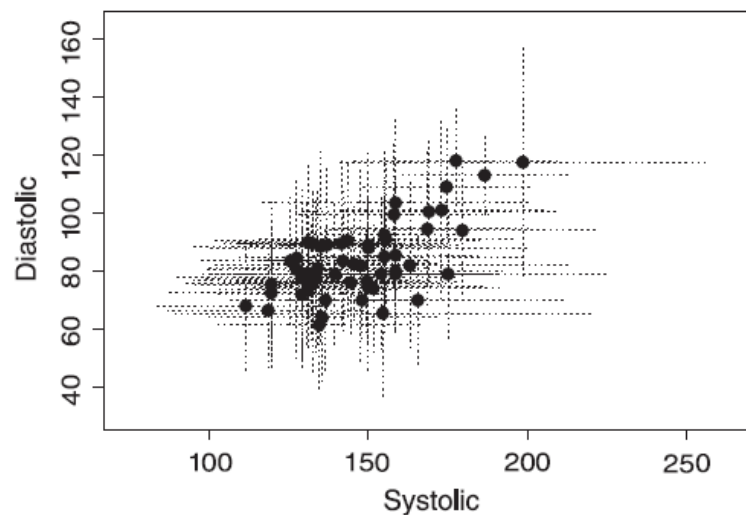


Figure 1.1. Interval value data visualisation example

Some attempts have been made to represent intervals, as shown in Figure 1.4 for interval time series or in Figure 1.1 for a type of scatter plot for two different interval valued data variables. This project will include some of these existing plots and create some new interval valued data representations for further visual analysis and understanding.

1.1.2 Regression

Regression is one area where research has advanced recently. Intervals, unlike crisp data, have many significant features (centre, radius, lower limit and upper limit) making classical regression methods complicated to apply. The first ever basic methods created for interval regression were by Billard and Diday in 2000 [2] and in 2002[3]. Later they were picked up and improved by Neto and Carvalho in 2008[4] and 2010[5].

Many of these methods, specially the earliest ones, used crisp-value regression analysis applied to different parameters of the interval; the most recent ones however, like Sinova et al in 2012[6] or Souza et al in a paper in 2017[7] try to go deeper and use techniques like the arithmetic of intervals or its parametrization.

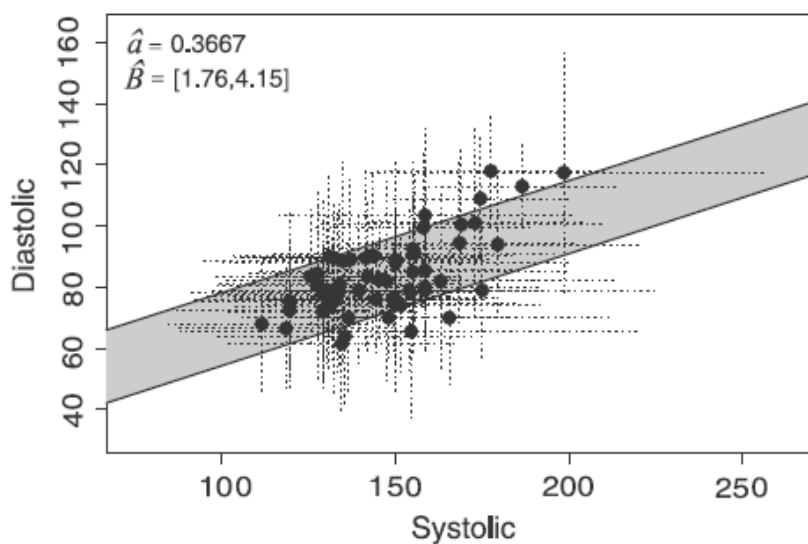


Figure 1.2. Interval value regression model example

For this project, a wide range of regression methods will be tested and included in the tool, some of the original ones from Billard and Diday [2] and Neto and Carvalho in 2008[4] and some more recent and

advanced methods that have been proposed in the last couple of years [7, 14, 15].

Under MATLAB, there is no functions for interval valued data regression, the only similar package found is the one in R developed by Lima Neto et al in 2016 [13] that includes some of the methods of this tool, mainly the earlier ones from Billard and Diday [2] and [3] and Neto and Carvalho [4] and [5].

1.1.3 Interval time series (ITS)

Interval time series are probably the most researched area of interval valued data. Applications in the business and finance fields, where significant data like stock exchange values or currency fluctuation can be represented as interval time series, are where most of the interest has been drawn from.

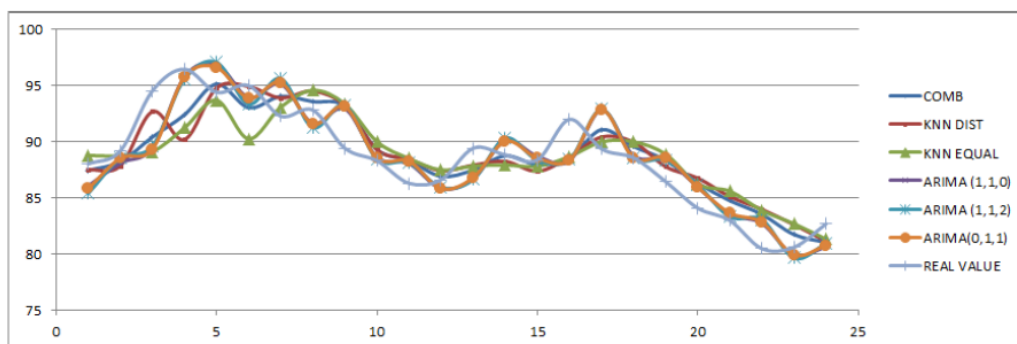


Figure 1.3. Interval time series forecasting example

For interval time series, even though it has been the most researched area of interval valued data, most of its analysis has focused on methods for forecasting like those found in Arroyo 2011[8] or master thesis like Morell in 2012[9]. Forecasting is the most useful analysis for the typical finance applications of Interval time series.

This project will not focus on forecasting, as it is a very broad and covered topic, and more on Interval time series representation and other analysis to be included in the tool.

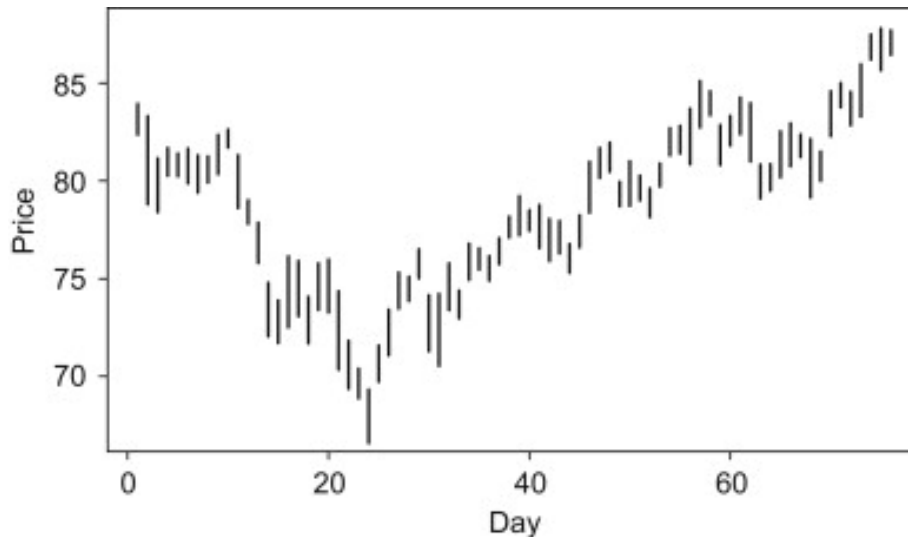


Figure 1.4. Interval time series visualisation example

1.1.4 Intervals in MATLAB

MATLAB has very little implemented on interval value data analysis. In 1999 a toolbox (INTLAB) was implemented with interval arithmetic[11] but that is as long as MATLAB has gone so far regarding interval valued data analysis and computation. Unlike R, where other libraries have been created with regression or visualization functions exist, MATLAB still has a long way to go regarding Interval Analysis giving room for this project to work with this need.

1.2 Motivation

As explained before, interval valued data has existed for a very long time, but its analysis is very recent and it is not completely extended and agreed upon. There are many alternative solutions to the same problem

and new issues arising from the exponentially increased amount of data available to us in today's world. The rapidly growing demand of data mining and data analysis will increase the need for tools to analyse all kinds of data.

On top of that, MATLAB, one of the most powerful and used tools for mathematical software, has very little tools and libraries for interval valued data analysis, operation and graphical representation.

With all this in mind, a project that tries to recollect and study the existing alternatives of interval valued data analysis and representation and creates a unifying tool in MATLAB to work with interval valued data can be very interesting and useful for future works.

1.3 Objectives

The final objective for this project is the completion of a MATLAB tool that can allow a user to upload a data set of interval-valued variables and help him or her analyse and work with them. For this ultimate objective, the following partial objectives have to be reached:

- Creation of different independent functions in MATLAB for the different functionalities such as visualisation, regression, sorting, selecting... etc
- Merge of all the different functions on a MATLAB tool that allows that various tests and analysis can be made to the same data set.
- Creation of a user-friendly interface to manage and run the created tool facilitating the file upload and the different options for the user.
- Test and trial of the interface with different methods and analysis of the different solutions given by the tool.

1.4 Work Methodology

As given by the objectives, and the previous descriptions, this project has two main parts with different methodologies:

1- Reading and researching of existing papers, articles, coding libraries and previous thesis which could provide useful tools for interval valued data analysis, regression, visualisation or measurement. Some will be recommended and provided by the director, others found through own research.

2- Coding and usage of MATLAB. This part is what would take most of the time: first the programming of the functions needed for the tool's different functionalities; once those functions are completed the creation of the tool itself which, because of its singularities, forced the modification and tuning of some details of the previously coded functions to match the requirements of the App designer.

1.5 Resources

For this project, the main resource to be used is going to be MATLAB. Many different tools from MATLAB were needed for the different parts of the project. The functions were coded in classic .m scripts, the tool, however, was created using the new MATLAB Graphical User Interface creator called MATLAB APP DESIGNER.

MATLAB App Designer is a tool introduced by MathWorks with MATLAB's 2016a version that modernises and simplifies the ability to create a GUI improving the previous tool for previous versions like MATLAB's own GUIDE.

MATLAB App Designer includes a "design view" as seen in Figure 1.5 where user can design the layout of the user interface, and

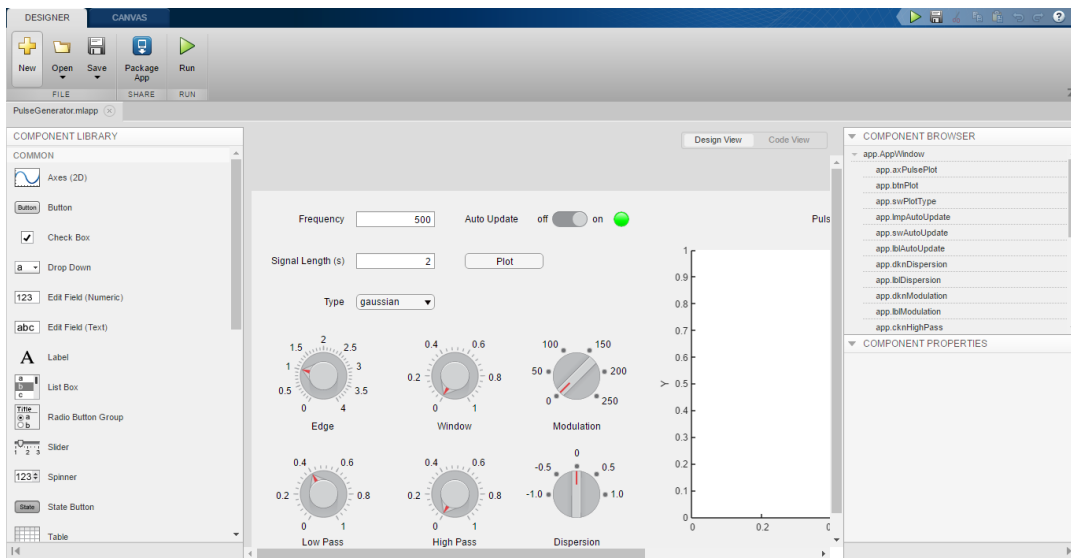


Figure 1.5. MATLAB App Designer design view

then it has a "code view" as seen in Figure 1.6 where the code of the layout is automatically generated according to positioning of the design view and where the user can program the code needed for the callback functions and execution of the app itself.

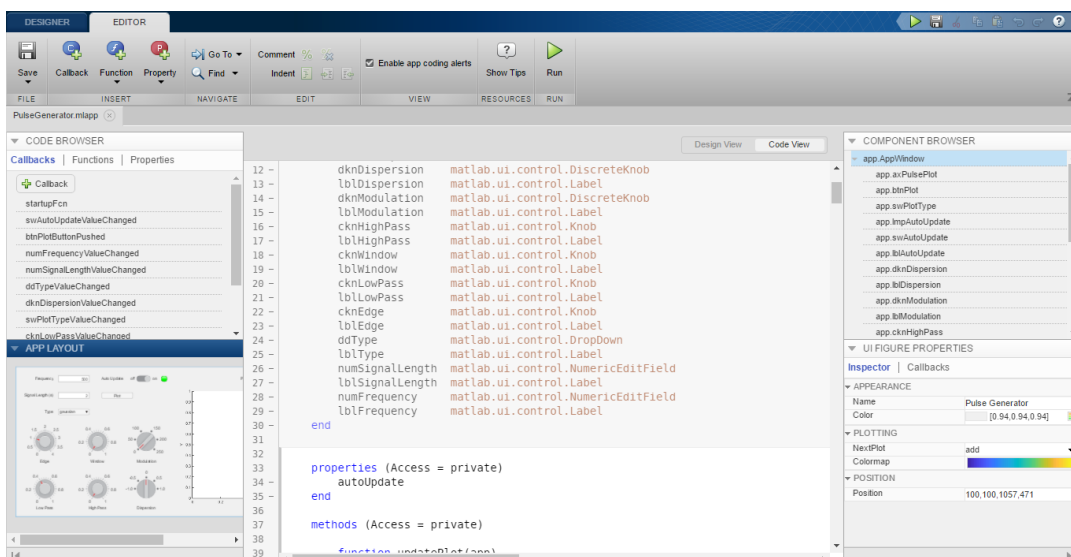


Figure 1.6. MATLAB App Designer code view

The final MATLAB tool used is the Application Compiler. This is a tool included in MATLAB as well that allows turning a MATLAB based

App like the one created for this project into a stand alone Application that can be installed in any computer with a .exe file even without having a MATLAB license.

The App Compiler, which may not seem as important, is very useful as it can help spread this tool to be used by anyone who wishes to use it even if they do not own a MATLAB license as it includes all the files and functions, even the ones created by the coder, needed for the design as seen in the bottom part of Figure 1.7.

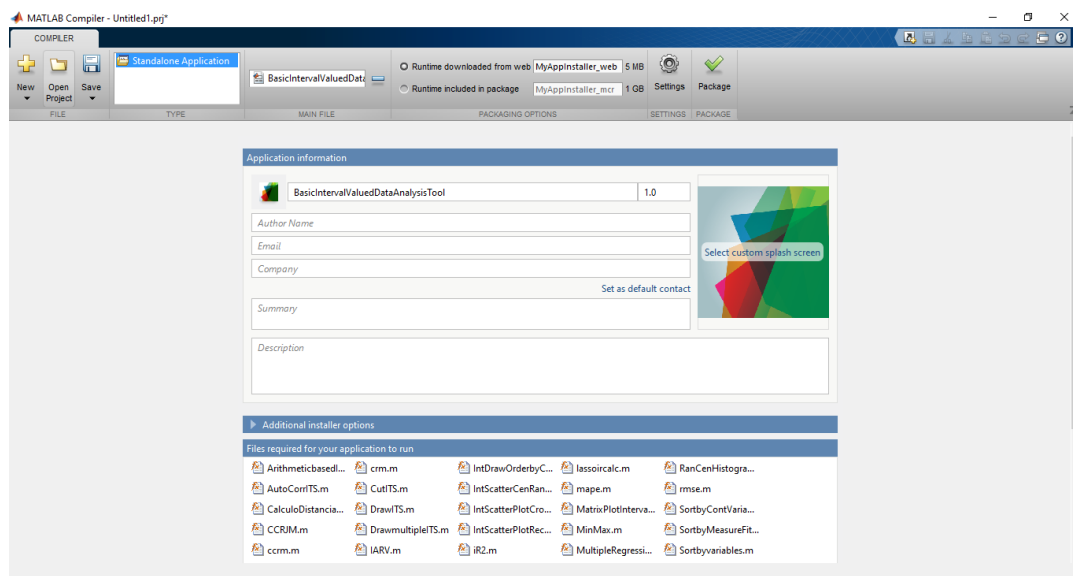


Figure 1.7. MATLAB App Compiler

On top of MATLAB, other software were used in a smaller way. Excel is the software needed for the datasets fed to the App so most of the functions coded take information of an .xlsx file and even write new ones in certain cases. 'R' was also used as some functions coded were already created in R programming and the effort was made to "translate" them to MATLAB for this tool.

For information and research, many interval-valued analysis articles will be used from different magazines such us *Computational Statistics*

and Data Analysis, Information Sciences or Knowledge-Based Systems as well as some some statistical in books like Rencher's[12] book.

For information regarding programming of the main functions and the APP Designer, MATLAB's own web page *www.mathworks.com* was used. This page includes clear documentation and examples on many types of functions needed for the project and functionalities that were included on the App. On top of that, its "MATLAB Answers" on its "community" section of the web page includes numerous answered and verified questions with solutions to coding and design problems that were extremely helpful during this project.

Chapter 2

Visualisation

As explained in the introduction, visualisation is one main area where interval valued data has to improve in. As intervals need many points to be represented (at least two) they become tough to represent specially if trying to include many variables, or some analytic plot.

In this project, different types of plots will be explored and included in the tool, some of them will be classic representations of interval valued data or ITS, whereas others will be creative alternatives proposed trying to emulate classic crisp plots, such as box plots or scatter plots, to represent different information aspects of and interval valued data variable.

2.1 Cross-sectional data visualisation

Cross-sectional interval valued data includes interval data from different entities (people, firms, regions.. etc) at the same point of time. Because of this reason the data has no order itself and allows rearrangement and, therefore, more freedom for visualisation.

2.1.1 Single variable

The first plot included is a simple representation of all the intervals in a variable as seen in Figure 2.1. This plot represents all intervals as a vertical line spanning the range of the entire interval. The Y axis gives the values of the intervals while all intervals are equally spaced and laid out along the X axis that does not represent anything.

This graph is the most intuitive way of representing intervals and, therefore, the typical way of representing interval valued data, whether cross-sectional or ITS, so far. It represents an issue when the amount of data is too big, the image can get too messy and difficult to visualise as all the lines turn into a big blur.

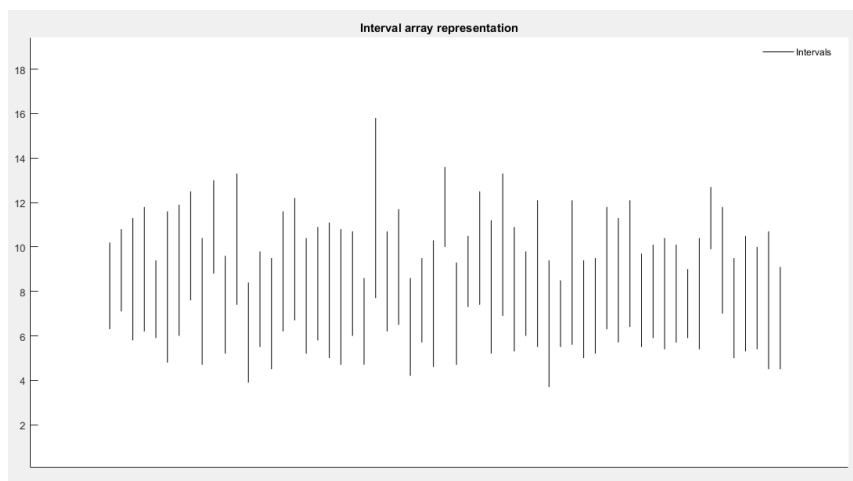


Figure 2.1. Graphical representation of all intervals on a single variable

The next graph proposed is a step further from just showing all the intervals and gives a basic visual analysis after some calculation. The visual inspiration is the classic "box plot" used with crisp data.

The "Main Intervals plot" (Figure 2.2) is a plot that includes only 5 intervals per variable:

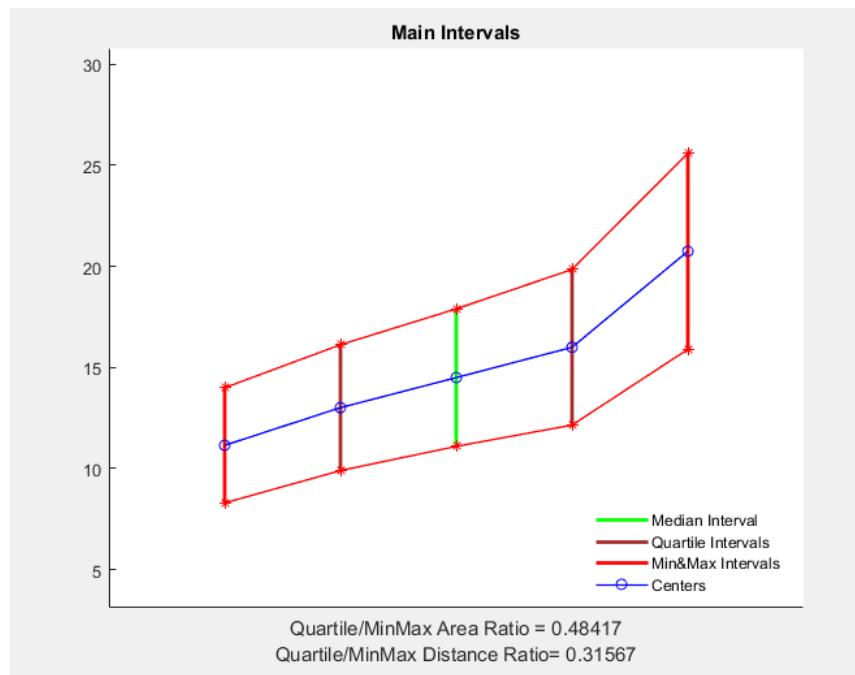


Figure 2.2. Main intervals plot

- **Minimum interval** This interval ranges from the minimum of the lower limits to the minimum of the upper limits, as shown by the left-most vertical red line.
- **Maximum interval** This interval ranges from the maximum of the lower limits to the maximum of the upper limits, as shown by the right-most vertical red line.
- **Median interval** This interval ranges from the median of the lower limits to the median of the upper limits as shown by the green line.
- **First quartile interval** This interval ranges from the first quartile of the lower limits to the first quartile of the upper limits, as shown by the left-most brown line.
- **Third quartile interval** This interval ranges from the third quartile of the lower limits to the third quartile of the upper limits, as shown by the right-most brown line.

Analysing the visual result of this plot, it has to be taken into account that the significant visual information from those intervals is not their horizontal separation as it would be in a traditional box plot with this orientation. The important indicator is, once again, the Y axis and, therefore how steep the MIP is and how vertically separated their intervals are.

On top of the visual intervals, this plot returns two parameters shown in the X axis. These parameters are, on some way, a measure of the dispersion of the data set. The first one is the Quartiles/MinMax area ratio which gives a ratio from 0 to 1 of the interquartile area divided by the total area comprised by the red lines. The other value is the Quartiles/MinMax distance ratio, which takes the average distance from the three distances worked with in this project: Hausdorff, Bertoluzza and Nucleus (detailed explanation of them in subsection 3.2.8) from the first quartile and the third quartile and divides it by the average distance from the maximum interval to the minimum. These parameters are a numeric value to be able to compare somehow multiple datasets and how spread their intervals are.

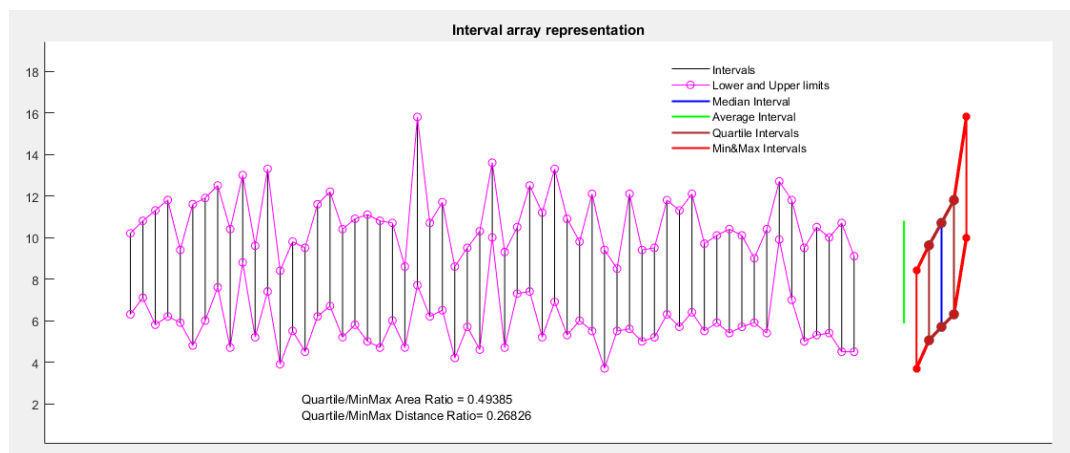


Figure 2.3. All intervals plus main intervals plot

The first two plots (Figure 2.1 and Figure 2.2) have the possibility to be combined into a greater plot as shown in Figure 2.3. This graph even adds the Average interval (from average of lower limits to average of upper limits, green line) on top of the main intervals and the dispersion values to give the full picture of the variable and give the user the chance to find what they are looking for. Once Again, if the amount of intervals is too big, this option maybe too messy to read, being the MIP the better choice to get an overall idea of the variable.

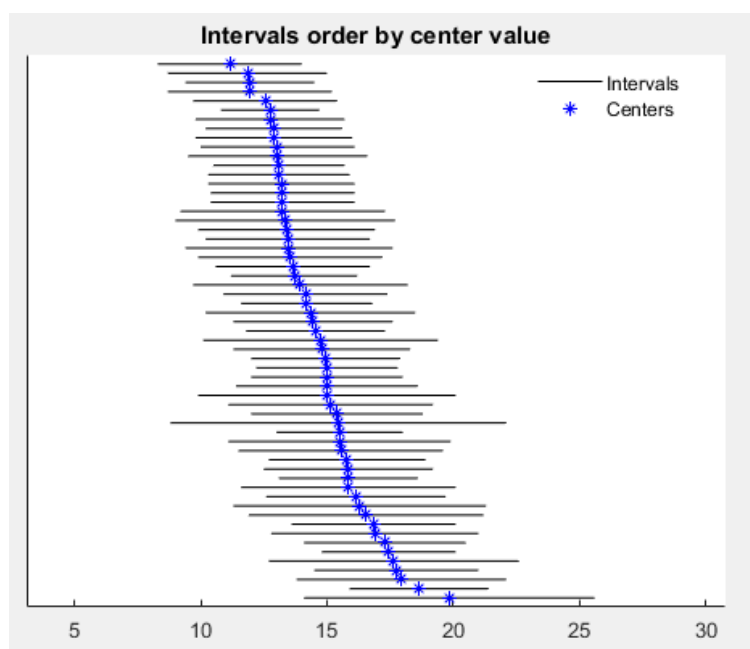


Figure 2.4. Intervals ordered by center value

The final two plots for a single interval valued data variable are those in Figure 2.4 and Figure 2.5.

The first one (Figure 2.4) shows all intervals ordered by their center value, which can be a better way to point out some outliers in the dataset. In the previous plots, with their roller-coaster-like shape, It is hard to identify possible outliers.

The final plot (Figure 2.5) is the interval version of the typical histogram used for a crisp variable. For interval valued data, each

interval at least needs 2 variables to be explained, either center and radius or minimum and maximum, with this in mind the 3D histogram is the only way to represent the frequency of intervals. With this graph, where the center of the intervals is represented in the X axis, the radius in the Y axis and the frequency on the Z axis.

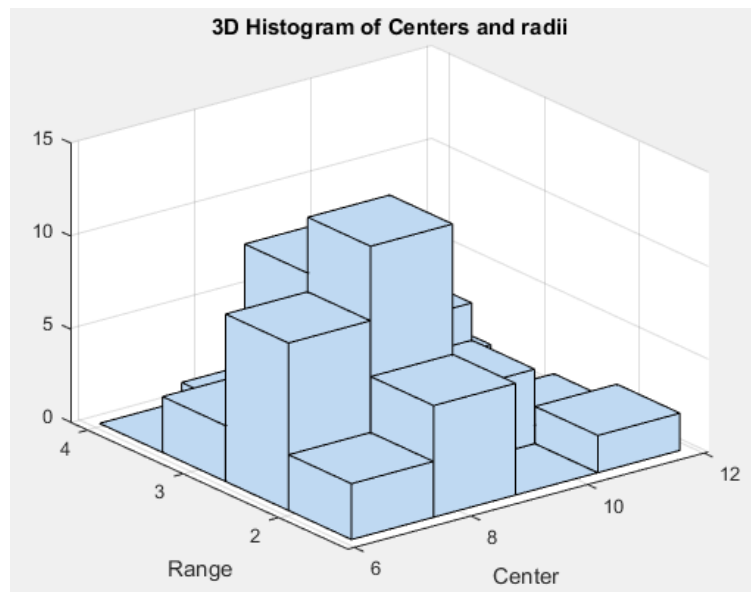


Figure 2.5. Three dimensional histogram of centers and radii

2.1.2 Multiple variables

Plotting many interval variables is not easy as they need many data points to be explained, however, there are a couple of ways two plot one interval valued data variable against another that are commonly used.

On Figure 2.6 one of the interval's versions of the scatter plot can be seen, Interval scatter plot using crosses. This plot emphasises the centers of the intervals of both variables, plotting the actual crisp scatter plot for them, and then crosses come out of each point to symbolise the radii of the intervals.

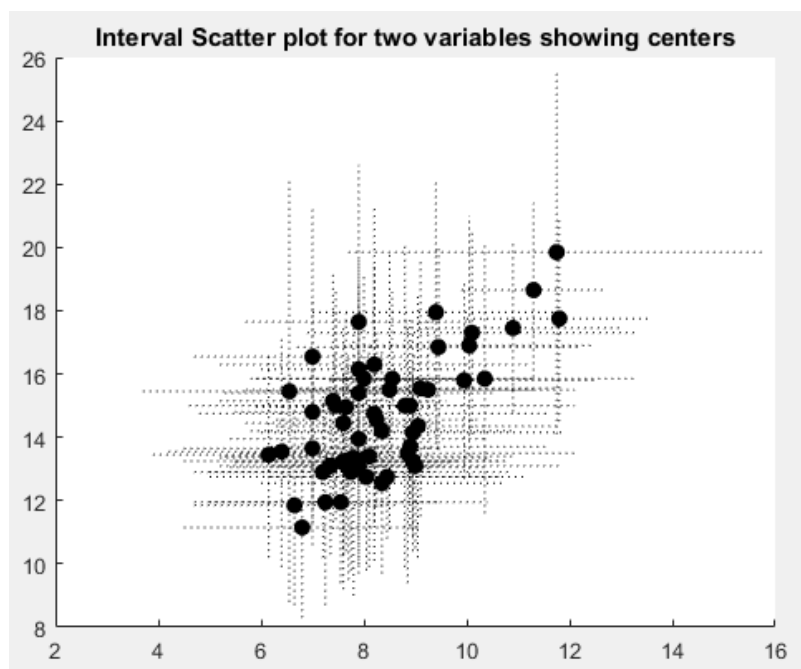


Figure 2.6. Interval scatter plot for two variables using crosses

On Figure 2.7 the second version of the interval scatter plot is shown. On this occasion, the centers are not shown and each pair of intervals is represented by a rectangle spanning the length of each interval in each dimension. As opposed to the interval scatter plot with crosses, this graphic focuses on the upper and lower limits of each pair intervals and the "area" that they cover.

Both of these graphs are very good for associating two interval variables and trying to find a possible correlation among them. These graphs are heavily used on many articles and will be a key component of the tool specially regarding regression, where these plots work best. On the use of one alternative or the other, it is completely up to the user's preferences. However, when the number of data intervals is not that big, the rectangles version shows a nice overview whereas when the number of data intervals is bigger, the rectangles get too crowded and messy to understand and the crosses version might seem the best

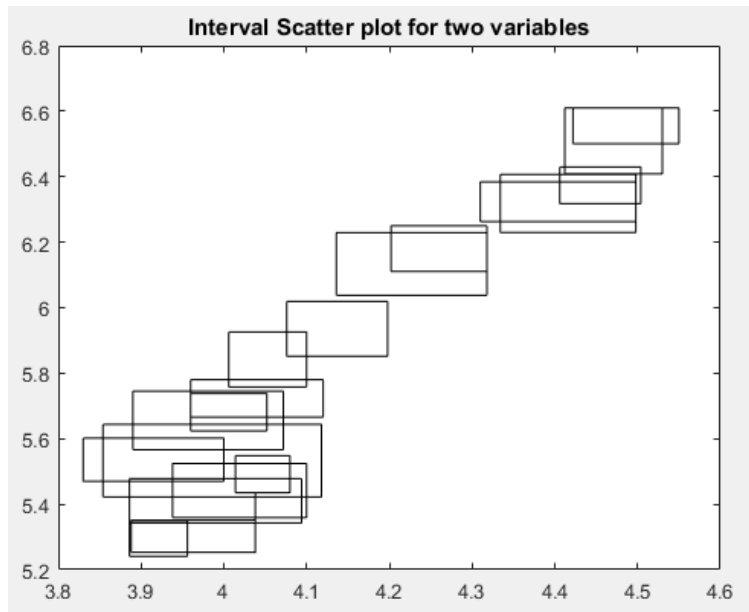


Figure 2.7. Interval scatter plot for two variables using rectangles

option. This may be due to the fact that the crosses version has more similarity with the actual scatter plot that we are used to.

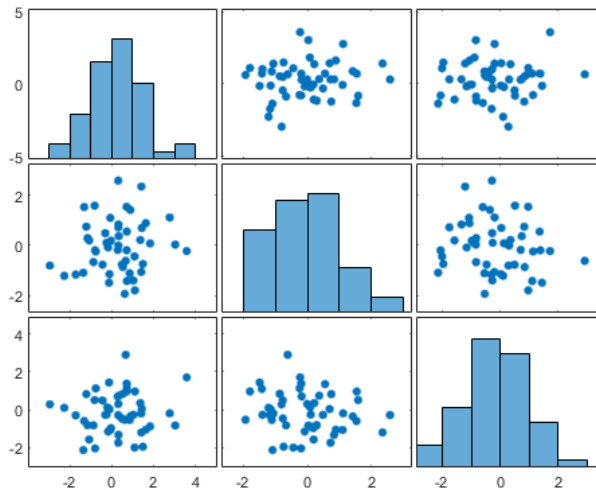


Figure 2.8. Classic crisp data matrix plot

So far all these graphs only represent one or two variables at the most, and no other plot is known for more variables, but taking as a reference the classic matrix plot (Figure 2.8) for multiple crisp variables

and the previous interval plots, a interval data matrix plot was included in the tool as seen in Figure 2.9.

Just like in the regular matrix plot, the non diagonal graphs represent scatter plots relating the two corresponding variables. In the interval version those scatter plots can either be the cross version or the rectangle version. Also, the diagonal plots, like in the regular matrix plot where it can be an histogram or a box plot, in the interval version can be comprised of any of the single variable plots explained earlier. In the case of Figure 2.9, it shows an interval regression plot with its crosses version and a diagonal made of Main Interval Plots for each one of the variables.

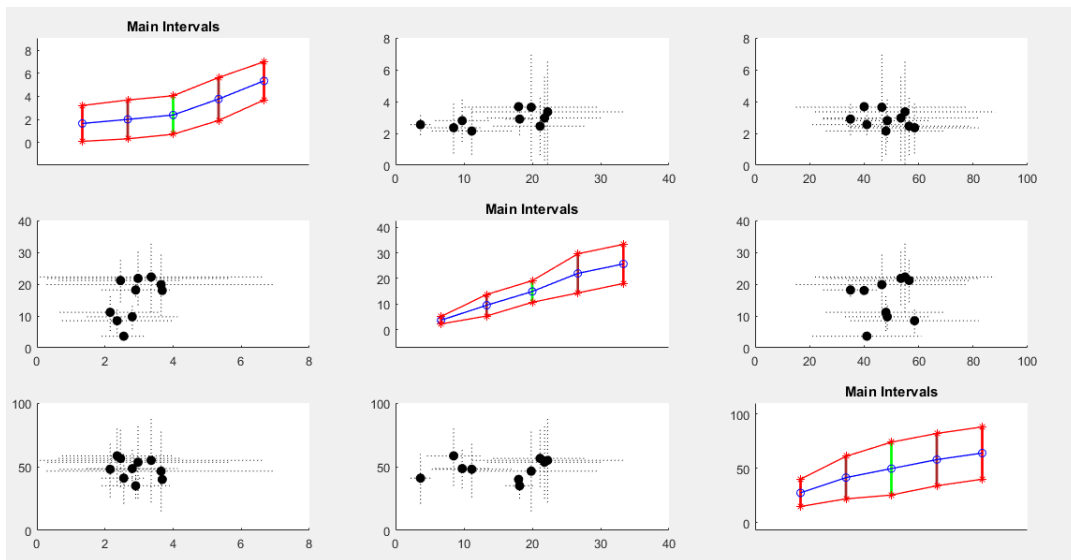


Figure 2.9. Interval matrix plot

2.2 ITS Visualization

Interval time series (ITS) are the most used interval valued data, so it is no surprise that there are also common ways for their visual representation. The most common one is that shown in Figure 1.3 for a single ITS or Figure 2.10 for multiple ITS. All intervals are shown

with straight vertical lines like the ones in Figure 2.1 but in this case the order is important as the X axis recovers its significance becoming the time axis to follow the ITS.

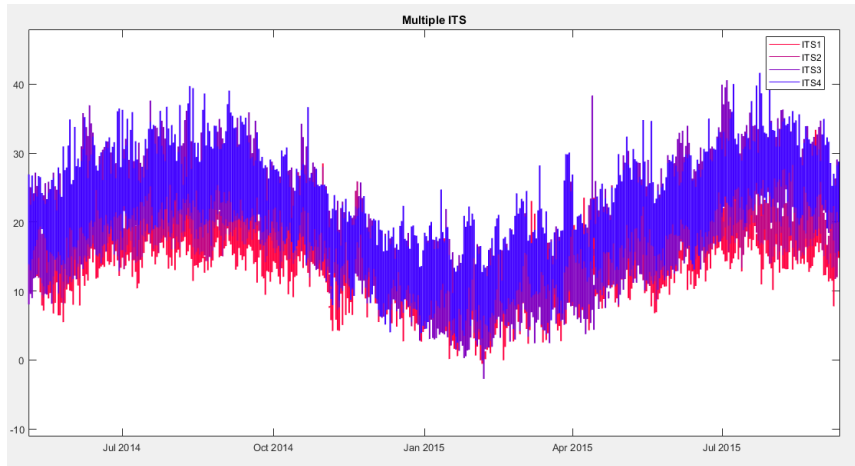


Figure 2.10. Multiple ITS of high and low temperatures in different cities during two years

Apart from this usual representation, two other plots were included in the tool to help understand and obtain more information out of ITS:

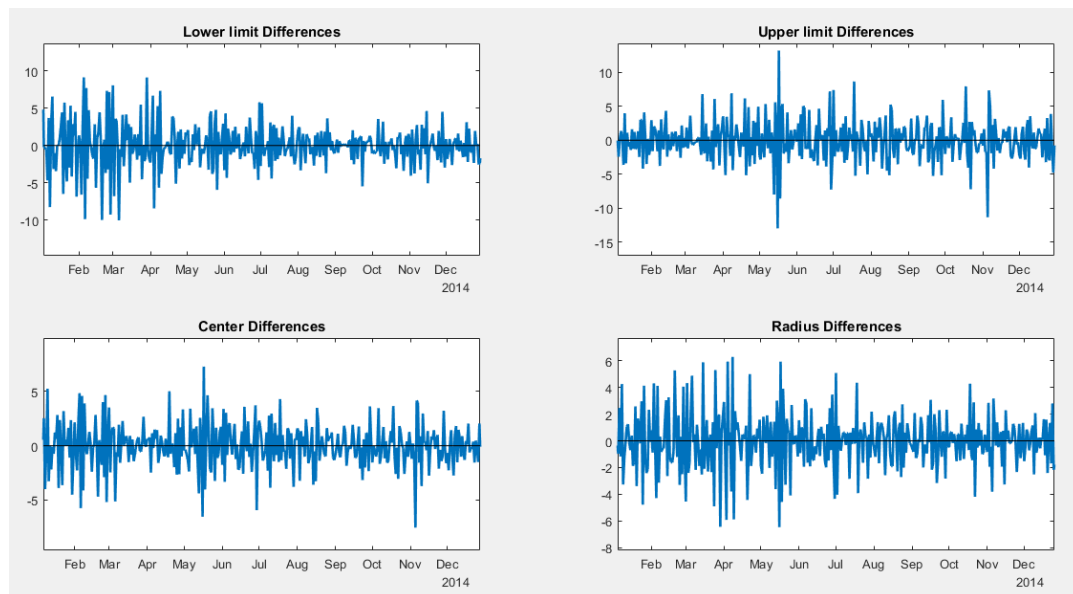


Figure 2.11. Variation of interval parameters from t_{k-1} to t_k

The first one is the one shown in Figure 2.11 in which 4 plots show the variation of each of the four main parameters of intervals (maximum, minimum, center and radius) from time t_{k-1} to t_k . This plot allows the option to see how volatile or stable the ITS, seeing if the change of its values is something mild or aggressive, if follows a trend or if it is random.

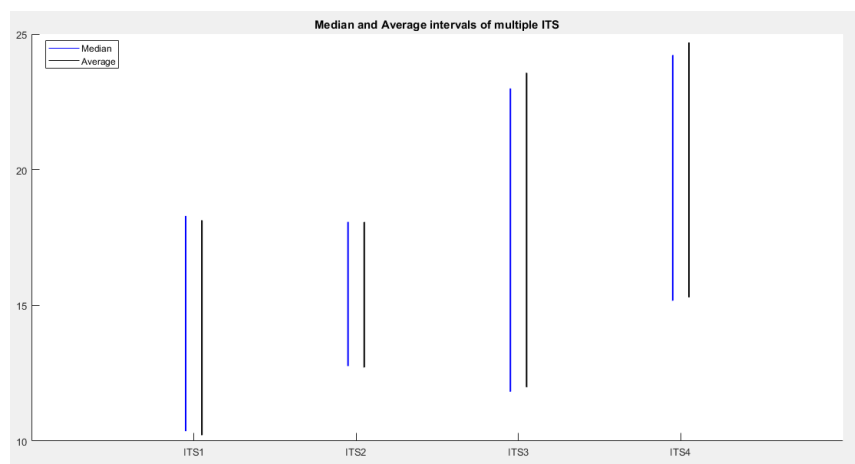


Figure 2.12. Average and median intervals for multiple ITS

The second plot is the one shown in Figure 2.12 that plots the average and median intervals of multiple ITS. This plot is useful to get a general idea of different ITS and be able to compare them when maybe there are too many data intervals. As an example, Figure 2.12 is the representation of the four ITS plotted in Figure 2.10 where the overlapping of ITS and the amount of data made it hard it to understand and compare the different cities. In this graph it is easy to see which cities were colder or hotter or which had more variation on its temperatures.

This graph not only works for ITS, it will be used in the tool to compare different interval valued data variables like the predicted values of the different regression models.

Chapter 3

Regression Analysis

FOR usual crisp data, regression is something that has been work with for a long time. Throughout the twentieth century many articles, books or manuals have already established a great foundation of regression methods for simple or multiple variables, one example of this is Rencher's *Methods of Multivariate Analysis* [12] published in 1995 which includes methods for simple or multiple, multivariate or univariate regression.

In interval valued data however, something as basic as a simple linear regression has barely been worked with in the last 15-20 years and no clear method has been established as the norm. For this project, the tool will only include simple and multiple univariate linear regression methods. This methods for interval regression have started to be developed very recently, all have been developed from 2000 on and most of them in this decade.

To talk about interval linear regression, 4 different paradigms should be first differentiated:

- Paradigm 1: $\beta_0 + \beta_1x_1 + \dots + \beta_nx_n = y$ where the independent variables x_n and the dependent variable y are all crisp data. This is the usual regression that has been worked with for decades.

- Paradigm 2: $\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n = [Y]$ where the independent variables x_n are crisp valued data and the dependent variable $[Y]$ is an interval valued variable.
- Paradigm 3: $\beta_0 + \beta_1 [X_1] + \dots + \beta_n [X_n] = y$ where the independent variables $[X_n]$ are interval valued variables and the dependent variable y is a crisp variable.
- Paradigm 4: $\beta_0 + \beta_1 [X_1] + \dots + \beta_n [x_n] = [Y]$ where both the independent and dependent variables are interval valued data.

For the scope of this project the linear regression methods used have been designed for paradigm 4 type problems only although they might work for other ones¹, it is not what they were created for and different models should be used.

3.1 Interval Linear Regression models

This section includes an overview of the 8 different methods of ILR included in the tool from oldest to most recently published. The bases of most of them is the application of classic crisp linear regression to different interval parameters.

3.1.1 Centre Method (CM)

The *Centre Method* was introduced by Billard and Diday in 2000 [2]. This method obtains a regression model from the centres of the intervals and applies such model to the upper and lower limits to obtain $[\hat{Y}]$.

¹Paradigm 2 is a special case of paradigm 4 where the input variables have radius=0 or lower limit=upper limit, therefore some of the methods will be able to work with Paradigm 2 type problems despite not being developed for them

Considering:

$$\begin{aligned} y_{li} &= \beta_0 + \beta_1 x_{l1} + \cdots + \beta_n x_{ln} + \epsilon_{li} \\ y_{ui} &= \beta_0 + \beta_1 x_{u1} + \cdots + \beta_n x_{un} + \epsilon_{ui} \end{aligned} \quad (3.1)$$

From 3.1 the *Sum of the squares of deviation* of this model will be specified as:

$$S_{cm} = \sum_{i=1}^n (\epsilon_{li} + \epsilon_{ui})^2 \quad (3.2)$$

Where ϵ_{li} and ϵ_{ui} from (3.1) are:

$$\begin{aligned} \epsilon_{li} &= y_{li} - \beta_0 - \beta_1 x_{l1} - \cdots - \beta_n x_{ln} \\ \epsilon_{ui} &= y_{ui} - \beta_0 - \beta_1 x_{u1} - \cdots - \beta_n x_{un} \end{aligned} \quad (3.3)$$

With a linear regression of the centres of $[Y]$ and $[X_1, X_2, \dots, X_n]$, it is possible to obtain the values of $\beta_0, \beta_1, \dots, \beta_n$ that minimise S from (3.2).

The resulting values $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n)$ will predict \hat{Y}_{CM} as follows:

$$\begin{aligned} \hat{y}_{li} &= \hat{\beta}_0 + \hat{\beta}_1 x_{l1} + \cdots + \hat{\beta}_n x_{ln} \\ \hat{y}_{ui} &= \hat{\beta}_0 + \hat{\beta}_1 x_{u1} + \cdots + \hat{\beta}_n x_{un} \end{aligned} \quad (3.4)$$

On Figure 3.1 an example of a simple interval linear regression is shown. The red line indicates the fit line of the regression performed for the centres of $[Y]$ and $[X]$. This fit would be used for the upper and lower limit values as explained in (3.4) to obtain \hat{Y}_{cm} .

This model was the first one and because of that it was still very rudimentary and not taking into account the range part of the intervals. Its prediction is not very precise as it will be seen later on this project.

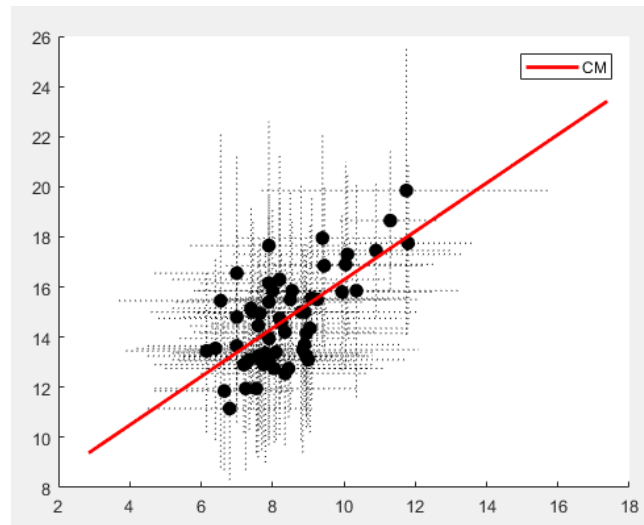


Figure 3.1. Centre Method regression model example

Despite that this method laid the foundation and the idea for further methods that came after it with similar approach.

3.1.2 MinMax method

The *MinMax* method was introduced by Billard and Diday in 2002 [3] as an improved alternative to the *Centre Method* developed by the same authors a couple year earlier. This method states the fact of using different parameters to obtain the lower and upper limits of $[\hat{Y}]$. This is the same as supposing independence between the values of the upper and lower bounds and treating them like they were different variables.

In this occasion the linear regression relationship works as follows:

$$\begin{aligned} y_{li} &= \beta_0^l + \beta_1^l x_{l1} + \cdots + \beta_n^l x_{ln} + \epsilon_{li} \\ y_{ui} &= \beta_0^u + \beta_1^u x_{u1} + \cdots + \beta_n^u x_{un} + \epsilon_{ui} \end{aligned} \quad (3.5)$$

From 3.5, once again, the *Sum of the squares of deviation* of this model will be specified as:

$$S_{minmax} = \sum_{i=1}^n (\epsilon_{li})^2 + \sum_{i=1}^n (\epsilon_{ui})^2 \quad (3.6)$$

Where ϵ_{li} and ϵ_{ui} from (3.5) are:

$$\begin{aligned} \epsilon_{li} &= y_{li} - \beta_0^l - \beta_1^l x_{l1} - \dots - \beta_n^l x_{ln} \\ \epsilon_{ui} &= y_{ui} - \beta_0^u - \beta_1^u x_{u1} - \dots - \beta_n^u x_{un} \end{aligned} \quad (3.7)$$

By minimising (3.6) taking into account the upper limits $[Y_u]$ and $[X_{u1}, X_{u2}, \dots, X_{un}]$ on a separate manner than the lower limits $[Y_l]$ and $[X_{l1}, X_{l2}, \dots, X_{ln}]$ by two independent linear regressions, it is possible to obtain the values of $\beta_0^l, \beta_1^l, \dots, \beta_n^l$ and $\beta_0^u, \beta_1^u, \dots, \beta_n^u$ that minimise S_{minmax} .

The resulting values $\hat{\beta}^l = (\hat{\beta}_0^l, \hat{\beta}_1^l, \dots, \hat{\beta}_n^l)$ and $\hat{\beta}^u = (\hat{\beta}_0^u, \hat{\beta}_1^u, \dots, \hat{\beta}_n^u)$ which will predict \hat{Y}_{MinMax} as follows:

$$\begin{aligned} \hat{y}_{li} &= \hat{\beta}_0^l + \hat{\beta}_1^l x_{l1} + \dots + \hat{\beta}_n^l x_{ln} \\ \hat{y}_{ui} &= \hat{\beta}_0^u + \hat{\beta}_1^u x_{u1} + \dots + \hat{\beta}_n^u x_{un} \end{aligned} \quad (3.8)$$

On Figure 3.2 an example of a simple interval linear regression using the Minmax method is shown. Both green lines indicate the fit line of the two different regressions performed by the upper limits and lower limits of $[X]$ to obtain $[\hat{Y}_L]$ and $[\hat{Y}_U]$.

This model is a clear upgrade from the Centre method as it takes into account the singularities that interval valued data have and by taking into account both upper and lower limits independently it makes a much more reliable estimation for $[\hat{Y}]$. It was the first real interval regression model that actually provided a decent estimation to work with and an example to look to by models developed later.

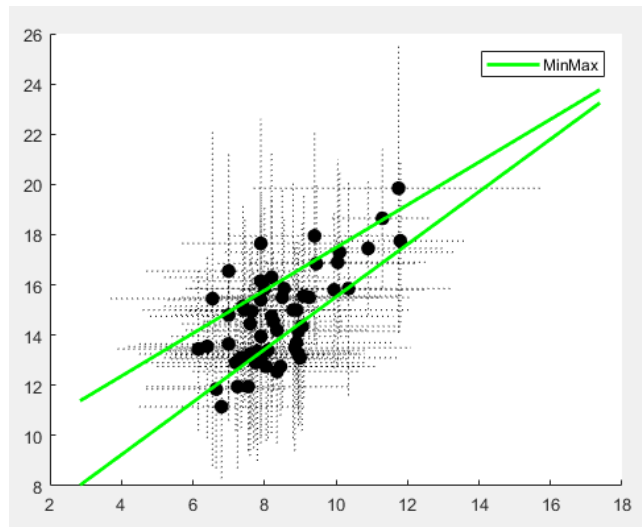


Figure 3.2. MinMax Method regression model example

3.1.3 Centre and Range Method (CRM)

The *Centre and Range Method* was introduced by Lima Neto and De Carvalho in 2008 [4] also as an improved alternative to the *Centre Method* developed Billard and Diday[2]. This method takes the idea of the Centre Method and decides to add the ranges or radii information to the model in order to improve the prediction performance. This method follows the idea of the Minmax method to add more information than the CM to meet the necessities of interval valued data complexity; in this case the centre and range.

In this occasion the interval W are not going to be represented by $[W_L W_U]$ but by $[W_c W_r]$ being $W_c = (W_L + W_U)/2$ and $W_r = (W_L - W_U)/2$.

For this model the linear regression relationship works as follows:

$$\begin{aligned} y_{ci} &= \beta_0^c + \beta_1^c x_{c1} + \cdots + \beta_n^c x_{cn} + \epsilon_{ci} \\ y_{ri} &= \beta_0^r + \beta_1^r x_{r1} + \cdots + \beta_n^r x_{rn} + \epsilon_{ri} \end{aligned} \quad (3.9)$$

From 3.9, the CRM method *Sum of the squares of deviation* of this model will be specified as:

$$S_{CRM} = \sum_{i=1}^n ((\epsilon_{ci})^2 + (\epsilon_{ri})^2) \quad (3.10)$$

Where ϵ_{ci} and ϵ_{ri} from (3.9) are:

$$\begin{aligned} \epsilon_{ci} &= y_{ci} - \beta_0^c - \beta_1^c x_{c1} - \dots - \beta_n^c x_{cn} \\ \epsilon_{ri} &= y_{ri} - \beta_0^r - \beta_1^r x_{r1} - \dots - \beta_n^r x_{rn} \end{aligned} \quad (3.11)$$

By minimising (3.10) taking into account the centres $[Y_c]$ and $[X_{c1}, X_{c2}, \dots, X_{cn}]$ like in the CM method and, on a separate manner, the radii $[Y_r]$ and $[X_{r1}, X_{r2}, \dots, X_{rn}]$ by two independent linear regressions, it is possible to obtain the values of $\beta_0^c, \beta_1^c, \dots, \beta_n^c$ and $\beta_0^r, \beta_1^r, \dots, \beta_n^r$ that minimise S_{CRM} .

The resulting values $\hat{\beta}^c = (\hat{\beta}_0^c, \hat{\beta}_1^c, \dots, \hat{\beta}_n^c)$ and $\hat{\beta}^r = (\hat{\beta}_0^r, \hat{\beta}_1^r, \dots, \hat{\beta}_n^r)$ will predict \hat{Y}_{CRM} as follows:

$$\begin{aligned} \hat{y}_{ci} &= \hat{\beta}_0^c + \hat{\beta}_1^c x_{c1} + \dots + \hat{\beta}_n^c x_{cn} \\ \hat{y}_{ri} &= \hat{\beta}_0^r + \hat{\beta}_1^r x_{r1} + \dots + \hat{\beta}_n^r x_{rn} \end{aligned} \quad (3.12)$$

On Figure 3.3 an example of a simple interval linear regression using the CRM method is shown. This method's fit line cannot be represented in a single interval scatter plot so it has been decided to show two regular scatter plots, the top one of the centres and the bottom one with the radii with both fit lines in red for the CRM method represented.

This model uses the same idea as the MinMax method to include more information to the CM method to improve its accuracy. It produces a similar performance as that of the MinMax method and obviously

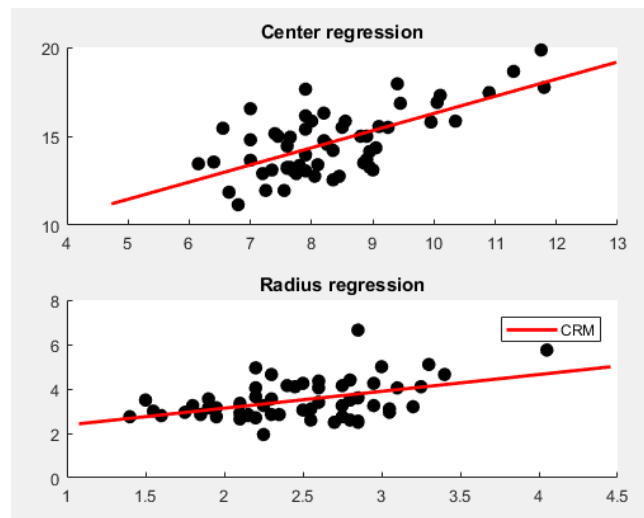


Figure 3.3. CRM method regression model example

improves that of the CM method adding the extra information needed for intervals.

3.1.4 Constrained Centre and Range Method (CCRM)

The *Constrained Centre and Range Method* was introduced by Lima Neto and De Carvalho in 2010 [5] as an improved version of their own CRM presented a couple years before [4]. The CCRM Method uses the same principle of the CRM of using the centre and radii of $[X_1, \dots, X_n]$ to predict $[\hat{Y}]$ but it includes a needed restriction that the radii cannot be negative, something that was possible with the CRM and that makes no mathematical sense.

The linear regression relationship is the same as the CRM method's one (3.9) but the *Sum of the squares of deviation* of this model will be specified as:

$$S_{CCRM} = \sum_{i=1}^n ((\epsilon_{ci})^2 + (\epsilon_{ri})^2) \quad (3.13)$$

$$s.t \quad \beta_i^r \geq 0, \quad i = 0, \dots, p.$$

By formalising (3.13) taking into account (3.11) in matrix notation the resulting optimisation problem is:

$$\min \quad \| Y_c - X_c \beta^c \|^2 + \| Y_r - X_r \beta^r \|^2 \quad (3.14)$$

$$s.t \quad \beta^r \geq 0.$$

The optimal values of β^c can be found by solving an ordinary regression problem taking into account that the second norm of (3.14) does not depend on β^c . Therefore, β^c will be the same as CRM. Similarly, the first norm of (3.14) can be considered as a constant for β^r . The estimate of β^r is obtained solving a constrained regression problem. This problem can be recognized as a particular non-negative least-squares (NNLS) problem. Therefore, CCRM consists of solving two separate regression problems.

The resulting values $\hat{\beta}^c$ and $\hat{\beta}^r$ will predict \hat{Y}_{CCRM} as follows:

$$\hat{y}_{ci} = \hat{\beta}_0^c + \hat{\beta}_1^c x_{c1} + \dots + \hat{\beta}_n^c x_{cn} \quad (3.15)$$

$$\hat{y}_{ri} = \hat{\beta}_1^r X_r$$

On Figure 3.4 an example of a simple interval linear regression using the CCRM method is shown compared to a CRM method. The blue line shows CCRM and the red line shows CRM. As expected, for the centre fit line, both methods are the same, whereas for the radius regression

there is a slight different derived from the constraint in the approach to the problem.

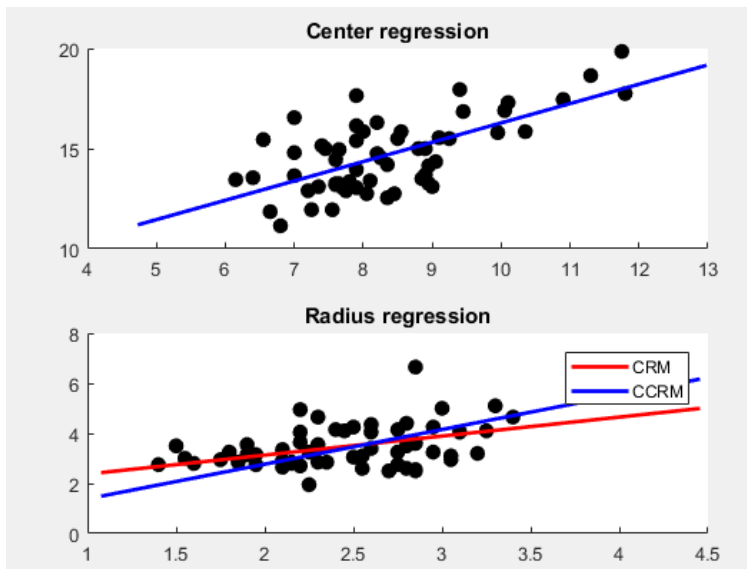


Figure 3.4. CCRM and CRM methods regression model comparison example

This model is a step forward from CRM as, not only does it provide a good prediction, but it accounts for cases that CRM or even MinMax would create an illogical result like intervals with negative radii or lower limits bigger than upper limits.

3.1.5 Arithmetic-based simple linear regression (ABSLR)

The *Arithmetic-based simple linear regression* was introduced by Sinova et al in 2012 [6]. This method is different from the previous ones exposed as it uses an arithmetic approach for the regression. It also differs on the fact that is the only method in the tool only for simple regression.

It should be noted that a linear model based on the usual interval arithmetic looks jointly at the centre and the radius of interval data (i.e., at each interval data as a whole) assuring that the interval arithmetic-based model is always well-defined. Instead, we could either examine

centre and radius separately or treating them as independent random variables. However, one cannot guarantee that a separate linear model for radius makes always sense like it has been talked about in previous models.

In a first approximation to the estimation problem under the linear model assumption, this problem accounts for finding a real value \hat{a} and an interval value \hat{B} such that $\hat{a}X + \hat{B}$ is as close as possible to the sample Y .

Considering the least squares problem of looking for \hat{a} and \hat{B} such that:

$$\frac{1}{n} \sum_{i=1}^n (d_{\theta}(Y, \hat{a}X + \hat{B})) \quad (3.16)$$

The algorithm to compute \hat{a} is as follows:

Step 1: Compute the sample estimates of the following moments of the real-valued random variables centres and radii:

– the means of the centres and radii of X and Y:

$$\overline{centx} = \frac{1}{n} \sum_{i=1}^n cent x_i, \quad \overline{radx} = \frac{1}{n} \sum_{i=1}^n rad x_i, \quad (3.17)$$

– the variances of the centres and radii of X and Y:

$$s_{centx}^2 = \frac{1}{n} \sum_{i=1}^n (cent x_i - \overline{centx}), \quad s_{radx}^2 = \frac{1}{n} \sum_{i=1}^n (rad x_i - \overline{radx}) \quad (3.18)$$

- the co-variances of the centres and radii of X and Y:

$$\begin{aligned} s_{centx,centy} &= \frac{1}{n} \sum_{i=1}^n (centx_i - \overline{centx})(centy_i - \overline{centy}) \\ s_{radx,rady} &= \frac{1}{n} \sum_{i=1}^n (radx_i - \overline{radx})(rady_i - \overline{rady}) \end{aligned} \quad (3.19)$$

Step 2: Compute the sample estimates of the following moments of random interval-valued set (for short RIS):

- the means of X and Y:

$$\overline{X} = [\overline{centx} - \overline{radx}, \overline{centx} + \overline{radx}], \quad \overline{Y} = [\overline{centy} - \overline{rady}, \overline{centy} + \overline{rady}] \quad (3.20)$$

- the θ -Fréchet variance of X:

$$\widehat{\sigma}_x^2 = s_{centx}^2 + \theta s_{radx}^2 \quad (3.21)$$

- the θ -covariance of X and Y, and -X and Y:

$$\widehat{\sigma}_{x,y} = s_{centx,centy} + \theta s_{radx,rady}, \quad \widehat{\sigma}_{-x,y} = -s_{centx,centy} + \theta s_{radx,rady} \quad (3.22)$$

Step 3: Compute the sample estimates \widehat{a}_0 and \widehat{a} for the available data:

- If all the x_i are real-valued, then $\widehat{a}_0 = \infty$.

- Otherwise:

$$\widehat{a}_0 = \min_{i: radx_i \neq 0} \frac{centx_i}{radx_i} \quad (3.23)$$

- if $\widehat{\sigma}_{x,y} < 0$ and $\widehat{\sigma}_{-x,y} < 0$, then, $\widehat{a} = 0$;

– if $\widehat{\sigma}_{x,y} \geq 0$ and $\widehat{\sigma}_{-x,y} \leq \widehat{\sigma}_{x,y}$, then:

$$\hat{a} = \min \left\{ \hat{a}_0, \frac{\widehat{\sigma}_{x,y}}{\widehat{\sigma}_x^2} \right\} \quad (3.24)$$

– if $\widehat{\sigma}_{-x,y} \geq 0$ and $\widehat{\sigma}_{x,y} \leq \widehat{\sigma}_{-x,y}$, then:

$$\hat{a} = \min \left\{ \hat{a}_0, \frac{\widehat{\sigma}_{-x,y}}{\widehat{\sigma}_x^2} \right\} \quad (3.25)$$

Once \hat{a} is obtained, the interval \hat{B} is obtained as:

$$\hat{B} = Y - \hat{a}X \quad (3.26)$$

The resulting values \hat{B} and \hat{a} will predict \hat{Y}_{ABSLR} as follows:

$$\hat{Y} = \hat{a}X + \hat{Y} \quad (3.27)$$

On Figure 3.5 an example of a simple interval linear regression using the ABSLR method is shown. The blue lines show the upper and lower bounds of the \hat{Y}_{ABSLR} . Compared to MinMax method, on this occasion upper and lower line fits have the same slope as \hat{a} is a Real value for both limits.

This model has a different approach from all the previous ones explained, the result obtained will be analysed latter compared to the other methods. The choice of θ in the tool has been that of 0.01 for all cases as the change of θ doesn't impact much the result.

3.1.6 Linear regression based on Lasso technique

The *Linear regression based on Lasso technique* was introduced by Paolo Giordani in 2015 [15]. this method may resemble the CCRM

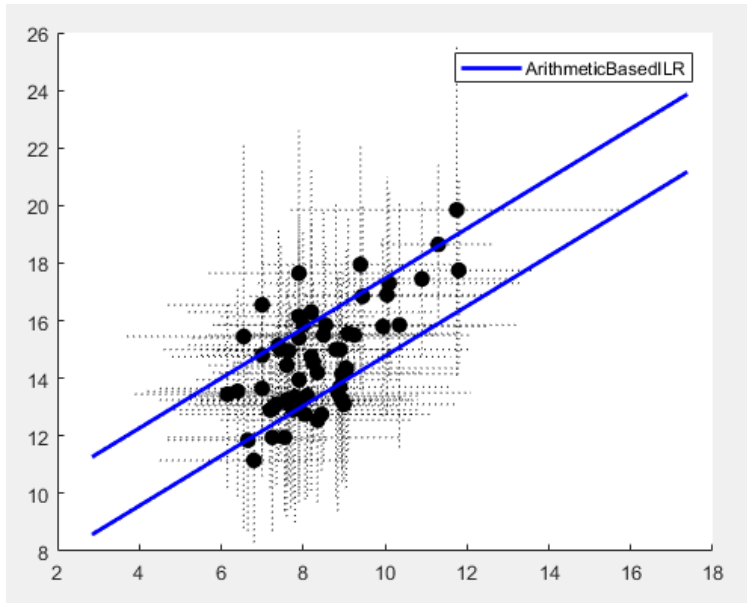


Figure 3.5. ABSLR method regression model example

approach because the problem is addressed as an optimisation problem involving the constrained minimisation of an objective function. It will attempt to seek a common set of regression coefficients for the midpoint and the radius. This will be done by adding specific regression coefficients for the radii in such a way to cope properly with all those situations in which the slope is different from the propagation of the imprecision.

For this model the linear regression relationship works as follows:

$$\begin{aligned} y_c &= b_c X_c + \epsilon_c \\ y_r &= b_r X_r + \epsilon_r = (b_c + b_a) X_r + \epsilon_r \end{aligned} \quad (3.28)$$

b_c and b_r are the vectors of the regression coefficients for the midpoint and radius models, respectively, where $b_r = b_c + b_a$ being b_a the vector of the additive coefficients. Therefore, the coefficients of the radius model b_r are equal to those of the midpoint model b_c up to the additive coefficients b_a .

The minimisation problem would be:

$$\min_{b_c, b_r} \| y_c - X_c b_c \|^2 + \theta \| y_r - X_c(b_c + b_a) \|^2 \quad (3.29)$$

A reasonable value for $\theta = \frac{1}{3}$ for further information see Trutschnig et al [16].

The loss function in (3.29) requires some constraints in order to guarantee that the estimated radii are non-negative and that b_a is as small as possible:

$$X_r(b_c + b_a) \geq 0 \quad (3.30)$$

The latter requirement can be managed using the Lasso technique. The Lasso, which is the acronym of Least Absolute Shrinkage and Selection Operator, is a method for estimation in regression with the objective of shrinking some regression coefficients and setting some others to 0. This is done by minimising the residual sum of squares with the constraint that the sum of the absolute values of the regression coefficients is smaller than a threshold. Resulting the lasso based problem in:

$$\begin{aligned} \min_{b_c, b_r} \| y_c - X_c b_c \|^2 + \theta \| y_r - X_c(b_c + b_a) \|^2 \\ \text{s.t. } X_r(b_c + b_a) \geq 0 \end{aligned} \quad (3.31)$$

The resulting model for \hat{Y}_{lasso} is such that:

$$\begin{aligned} \hat{y}_c &= b_c X_c \\ \hat{y}_r &= (b_c + b_a) X_r \end{aligned} \quad (3.32)$$

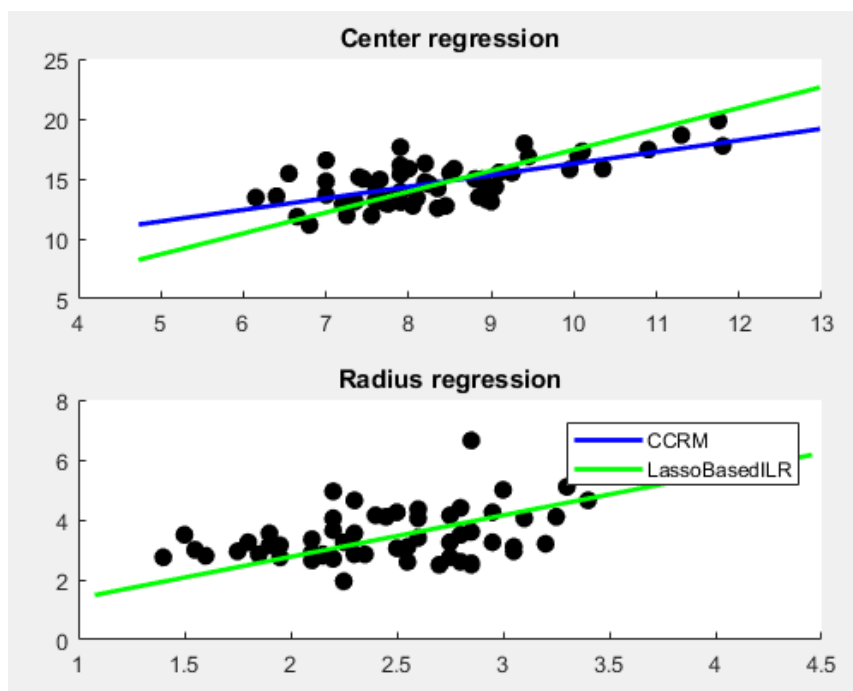


Figure 3.6. Lasso and CCRM methods regression model comparison example

On Figure 3.6 an example of a simple interval linear regression using the Lasso method is shown compared to the CCRM. The blue lines show the CCRM and the green lines show Lasso method. As expected the radius fit line is the same for both, as they use the same non-negative constraint to achieve it. The centre one, however, is different as the problem tries to minimise the difference between the b_c and b_r .

3.1.7 Constrained centre and range joint method (CCRJM)

The *Constrained centre and range joint method* was introduced by Hao and Guo in 2017 [14]. This method is a step forward from the CCRM method, it looks for a coefficient for the centre and the radius like CCRM does but does it jointly, it uses the radii values for the centre model and the centres for the radius model. It goes a step ahead because not only does it account for the fact that interval valued

data have multiple components like the previous methods, but it also acknowledges the relationship among those components.

Considering the linear regression relationship to be:

$$\begin{aligned} Y_c &= X\beta^c + \epsilon_c \\ Y_r &= X\beta^r + \epsilon_r \end{aligned} \quad (3.33)$$

However, in this case unlike 3.9, X is not only the centres or the radii, $X = [1 \| X^c \| X^r]$ where $X^c = \begin{bmatrix} x_{11}^c & \dots & x_{1p}^c \\ \vdots & \ddots & \vdots \\ x_{n1}^c & \dots & x_{np}^c \end{bmatrix}$ with all the centre values and $X^r = \begin{bmatrix} x_{11}^r & \dots & x_{1p}^r \\ \vdots & \ddots & \vdots \\ x_{n1}^r & \dots & x_{np}^r \end{bmatrix}$ with all the radii values.

Therefore, the vector β^c and β^r will have dimension $2p+1$ as they will need 1 coefficient for the centre and radius of each X variable.

Thus, the minimisation problem is given, once again by:

$$\min \sum_{i=1}^n \epsilon_{ci}^2 + \epsilon_{ri}^2 \quad (3.34)$$

With the non-negativity constraint, this time using the entire matrix X :

$$s.t. \quad X\beta^r \geq 0 \quad (3.35)$$

The resulting prediction for \hat{Y}_{CCRJM} is:

$$\begin{aligned} \hat{Y}_c &= X\beta^c \\ \hat{Y}_r &= X\beta^r \end{aligned} \quad (3.36)$$

Unfortunately for this method there is no graphical output for the fit line, the fact of using centre and radius value for both components turns even a simple interval valued linear regression into a multiple regression being unable to represent its result in a 2D plot. Its effectiveness will be later tested numerically with the other methods.

3.1.8 Parametrized Approach for Linear Regression of Interval Data

The *Parametrized Approach for Linear Regression of Interval Data* was introduced by Souza et al in 2017 [7]. Just like CCRJM, this method is a step forward from the MinMax method, it looks for a coefficient for the lower limit and the upper limit like MinMax method does but does it jointly, it uses the lower limit values for the upper limit model and the upper limits for the lower limit model. Like CCRJM It goes a step ahead because not only does it account for the fact that interval valued data have multiple components like the previous methods, but it also acknowledges the relationship among those components.

Considering the linear regression relationship to be:

$$\begin{aligned} Y_u &= X\beta^u + \epsilon_u \\ Y_l &= X\beta^l + \epsilon_l \end{aligned} \tag{3.37}$$

Where:

$$\begin{aligned}
X &= \begin{bmatrix} 1 & x_{11}^l & x_{11}^u & x_{21}^l & \dots & x_{p1}^l & x_{p1}^u \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{1n}^l & x_{1n}^u & x_{2n}^l & \dots & x_{pn}^l & x_{pn}^u \end{bmatrix} \\
\beta^l &= \left[\beta_0^l \quad \alpha_1^l \quad \omega_1^l \quad \alpha_2^l \quad \omega_2^l \quad \dots \quad \alpha_p^l \quad \omega_p^l \right]^T \\
\beta^u &= \left[\beta_0^u \quad \alpha_1^u \quad \omega_1^u \quad \alpha_2^u \quad \omega_2^u \quad \dots \quad \alpha_p^u \quad \omega_p^u \right]^T
\end{aligned}$$

Just like CCRJM, the vector β^l and β^u will have dimension $2p+1$ as they will need 1 coefficient for the lower limits and one for the upper limits of each X variable.

As this method takes into account all parameters for its model, the minimisation model is independent for obtaining β^u and β^l :

$$\begin{aligned}
S^l &= \sum_{i=1}^n (\epsilon_{li})^2 \\
S^u &= \sum_{i=1}^n (\epsilon_{ui})^2
\end{aligned} \tag{3.38}$$

The solution for β^l in the matrix mode is obtained as:

$$\beta^l = ((X)^T X)^{-1} (X)^T y^l \tag{3.39}$$

And for β^u :

$$\beta^u = ((X)^T X)^{-1} (X)^T y^u \tag{3.40}$$

The resulting prediction for \hat{Y}_{param} is:

$$\begin{aligned}\hat{Y}_l &= X\beta^l \\ \hat{Y}_u &= X\beta^u\end{aligned}\tag{3.41}$$

For this method, like in CCRJM there is no graphical output for the fit line, for the same reason: it turns even a simple interval valued linear regression into a multiple regression being unable to represent its result in a 2D plot. Its effectiveness will be later tested numerically with the other methods as well.

3.2 Goodness of fit test

For the goodness of fit test, a total of ten different parameters have been included in the tool. Some of them are interval oriented, some are transformation of classic crisp parameters into their interval version.

3.2.1 iR^2

The *coefficient of determination* (R^2) is a commonly used coefficient that determines a model's ability to predict results. For crisp data it is calculated as follows:

$$R^2 = \frac{\sigma_{X,Y}^2}{\sigma_X^2 \sigma_Y^2}\tag{3.42}$$

The result will be comprised between 0 and 1 and can be interpreted as a percentage, the closer to 1, the more X predicts Y. However, for this tool, its interval version was needed, for that matter the formula proposed by Ángela Blanco [18] was used:

$$iR^2 = \frac{\sigma_{centY}^2 R_{cent}^2 + \sigma_{radY}^2 R_{rad}^2}{\sigma_{centY}^2 + \theta \sigma_{radY}^2}\tag{3.43}$$

For this tool θ was taken as 1 for all methods to simplify and to turn iR^2 into a weighted average. The closer a model's iR^2 is to 1, the better its prediction is.

3.2.2 RMSE

The *root-mean-square error* (RMSE) is a commonly used coefficient that determines a model's difference between the estimation and the values observed. It is also a typical crisp data coefficient. For crisp data it is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (3.44)$$

The result is an absolute error, so the smaller it is the better the model is. For the tool, as RMSE is a crisp coefficient, the RMSE has been calculated for radius, centre, upper limit and lower limit (crisp values on their own) and the average of the four is returned as the RMSE value for the predicted interval.

3.2.3 MAPE

The *mean absolute percentage error* (MAPE) is another commonly used coefficient that determines a model's accuracy. It usually expresses accuracy as a percentage. For crisp data it is calculated as follows:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{y_i} \quad (3.45)$$

The result is also an error value, so the smaller it is the better the model is. For the tool, as MAPE is, once again, a crisp coefficient, the MAPE has been calculated for radius, centre, upper limit and lower

limit (crisp values on their own) and the average of the four is returned as the MAPE value for the predicted interval.

3.2.4 iARV

The *interval average relative variance* (iARV) is an interval valued data variable coefficient itself. As explained by C.Maté in 2015 [19] it compares the predictions of the model with the predictions given by average values of the interval variable. iARV is calculated as follows:

$$iARV = \frac{\sum_{i=1}^n (\hat{y}_i^l - y_i^l)^2 + \sum_{i=1}^n (\hat{y}_i^u - y_i^u)^2}{\sum_{i=1}^n (\bar{y}_i^l - y_i^l)^2 + \sum_{i=1}^n (\bar{y}_i^u - y_i^u)^2} \quad (3.46)$$

Lower values of iARV mean better estimations, it converges to $iARV = 0$ when the regression model is perfect.

3.2.5 iUTheil

The *interval Theil's U statistic* (iUTheil) is an interval valued data coefficient itself used only for ITS. As explained by C.Maté in 2015 [19] it compares the predictions of the model with naïve model.

iUTheil is calculated as follows:

$$iARV = \sqrt{\frac{\sum_{t=2}^n (\hat{y}_t^l - y_t^l)^2 + \sum_{t=2}^n (\hat{y}_t^u - y_t^u)^2}{\sum_{t=2}^n (y_t^l - y_{t-1}^l)^2 + \sum_{t=2}^n (y_t^u - y_{t-1}^u)^2}} \quad (3.47)$$

The U statistic compares the model to a random walk. If $iUTheil = 1$ it performs like a random walk, if $iUTheil > 1$ it performs worse than a random walk and if $iUTheil < 1$ it performs better, the lower it is, the better the prediction. Theil's U will only be included in the ITS part of the tool as it has no logical value for cross sectional variables.

3.2.6 Coverage Rate

The *Coverage Rate* (CR) is an interval coefficient. As explained by C.Maté in 2015 [19] the coverage rate is the average of all the ratios of the intersecting part of actual value intervals with the estimated ones divided by the length of the observed intervals. For further understanding better look at Figure 3.7. It is calculated as follows:

$$CR = \frac{1}{n} \sum_{i=1}^n \frac{w([\hat{y}_i^l] \cap [y_i^l])}{w([y_i^l])} \quad (3.48)$$

Being $w([\hat{y}_i^l] \cap [y_i^l])$ the distance of the intersection of the predicted interval and the observed value and $w([y_i^l])$ the distance of the observed value.

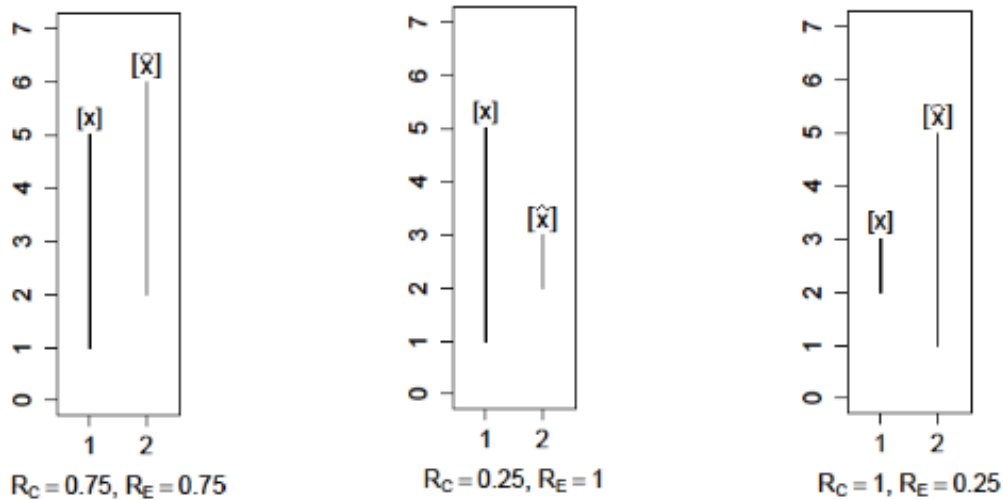


Figure 3.7. Illustration of the coverage and efficiency rates

3.2.7 Efficiency Rate

The *Efficiency Rate* (ER) is an interval coefficient. As explained by C.Maté in 2015 [19] the efficiency rate is the average of all the ratios of the intersecting part of actual value intervals with the estimated

ones divided by the length of the predicted intervals. For further understanding better look at Figure 3.7. It is calculated as follows:

$$CR = \frac{1}{n} \sum_{i=1}^n \frac{w([\hat{y}_i^l] \cap [y_i^l])}{w([\hat{y}_i^l])} \quad (3.49)$$

Being $w([\hat{y}_i^l] \cap [y_i^l])$ the distance of the intersection of the predicted interval and the observed value and $w([\hat{y}_i^l])$ the distance of the predicted value.

The closer these two values (ER and CR) are to 1, the better the prediction is, however as seen in Figure 3.7, one of the two rates equal to 1 might not mean a great prediction if the other value is very small.

3.2.8 Hausdorff Distance

The *Hausdorff distance* measures how far two subsets of the same metric space are from each other. For the interval method used in this tool, the formula is:

$$\text{Hausdorff dist} = \frac{1}{n} \sum_{i=1}^n | \hat{y}_i^c - y_i^c | + | \hat{y}_i^r - y_i^r | \quad (3.50)$$

This distance is an absolute value as well, therefore, the smaller the value, the better the model is.

3.2.9 Nucleus Distance

The *Nucleus distance* or Euclidean distance is the typical mathematical distance calculated by Pythagoras' theorem. For the interval method used in this tool, the formula used is:

$$\text{Nucleus dist} = \frac{1}{n} \sum_{i=1}^n \sqrt{(\hat{y}_i^c - y_i^c)^2 + (\hat{y}_i^r - y_i^r)^2} \quad (3.51)$$

This distance is an absolute value as well, therefore, the smaller the value, the better the model is.

3.2.10 Bertoluzza Distance

The *Bertoluzza distance* is a variation of the Nucleus distance proposed by Trutschnig et al [16] where the weight of the radius is smaller. Many studies have been developed since on this metric and it seems to be more appropriate for intervals or fuzzy data. For the interval method used in this tool, the formula is:

$$\text{Bertoluzza dist} = \frac{1}{n} \sum_{i=1}^n \sqrt{(\hat{y}_i^c - y_i^c)^2 + \theta(\hat{y}_i^r - y_i^r)} \quad (3.52)$$

Where $\theta = \frac{1}{3}$ as the standard value.

This distance is an absolute value as well, therefore, the smaller the value, the better the model is.

3.3 Testing and comparison of different methods

On this section the various regression methods explained on this chapter are going to be tested in a few datasets themselves. They are going to be tested with simple and multiple regression and using the goodness of fit tests explained in this chapter as well to determine the most recommended model.

3.3.1 Simple regression examples

The first example is classic dataset used on the articles of many of the methods explained. It is composed of 59 intervals of blood pressure values, X corresponds to the values of diastolic pressure during a day, and Y corresponds to the values of systolic pressure on the same day.

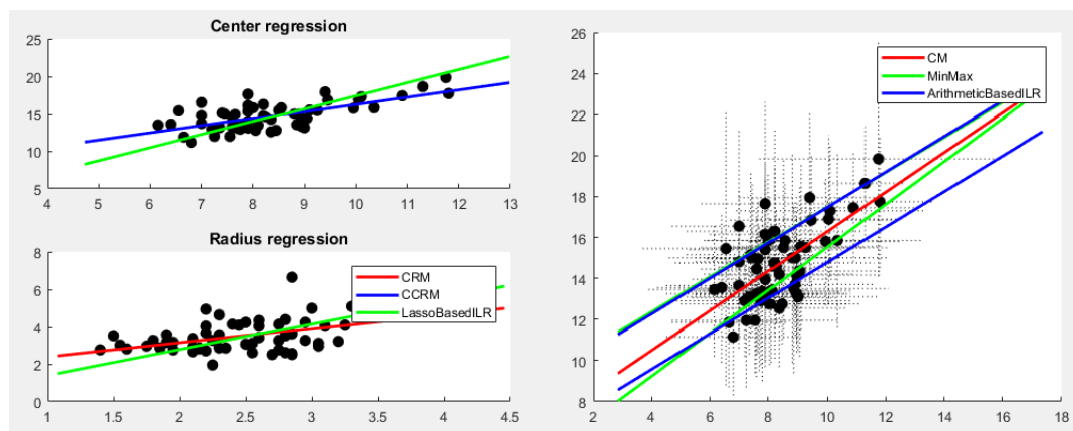


Figure 3.8. Graphical representation of all model fits

On Figure 3.8 all the graphical representation of the model line fits are shown, however this does not help a lot to determine which models are better.

On Figure 3.9 a comparison of all the predicted variable's median and average interval to Y 's median and average intervals (the left-most) is shown. In this graph more can be seen, like the fact that CM looks

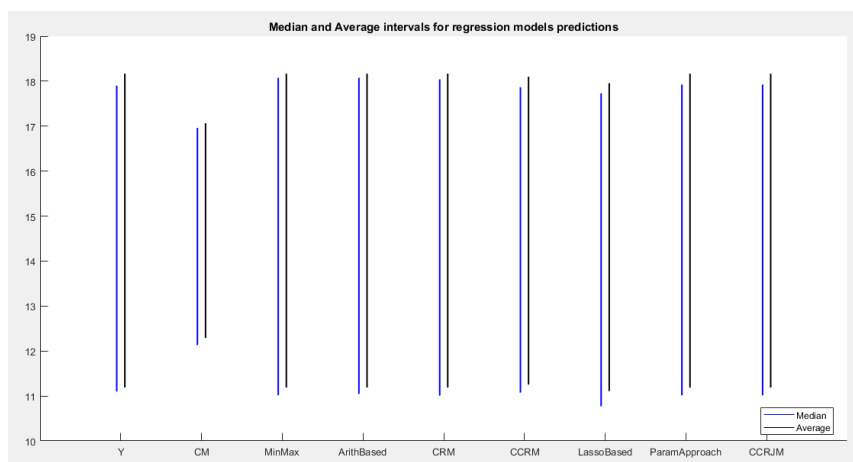


Figure 3.9. Median and average interval from Y and \hat{Y} of all the models

off, but the rest are all very similar to Y . Therefore, it is better to go to the actual numeric values to determine which is better.

Looking at Table 3.1 we can get an idea of the performance of each method. The first thing that stands out is that CM method performs the worst in every test shown except ER (as explained in ER section, it gives no information if the CR is bad, which is the case). This makes sense as its simplistic approach makes it hard to predict intervals as good as the other methods.

The second worst seems to be Lasso for every metric as well, its approach in favour of parsimony to reduce coefficients seems to affect the accuracy. The remaining 6 methods are all very close to each other and alternate positions depending on the metric and from best to worst barely varies more than 4-5%. The two best are CCRJM and ParamApproach as their similar detailed approach (in this example they seem to give the same result) seems to give a slight extra accuracy.

The second example to be tested is a weird one, it includes two different stock exchange values for 2 Spanish banks for 20 days. The weird thing about this dataset is that the intervals are very small in radius, in most cases the radius is barely 5% of the centre value.

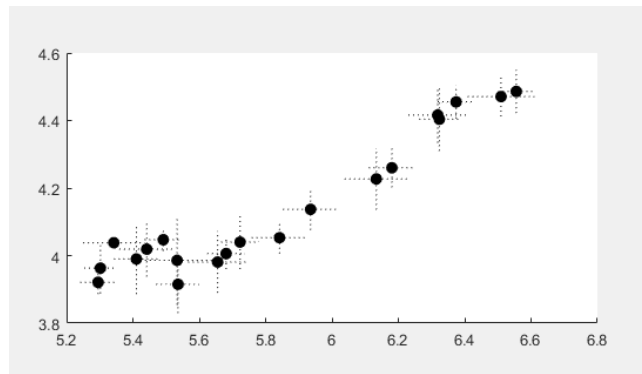


Figure 3.10. Intervals for second simple regression example

By looking at Table 3.2 the takeaways vary from the first example. For starters, CM now, as these intervals have very little value in their radii, performs in a similar level as the other methods. Also CCRJM, while still being one of the better ones, performs worse than ParamApproach as the first is based on the radius variable and in this example the radii are very small and volatile not helping the calculations. Lasso is still the worst in most of the cases, the lack of radius information also hurts it. The remaining methods, once again perform very similar to each other and vary their position depending on the metric used.

Regression Models								
	CM	MinMax	ArithBased	CRM	CCRM	LassoBased	ParamApproach	CCRJM
IR2	0.0929	0.3948	0.3885	0.3930	0.3665	0.1307	0.3986	0.3986
RMSE	1.6262	1.2887	1.2972	1.2917	1.3233	1.5452	1.2848	1.2848
MAPE	9.5381	7.6785	7.7274	7.7030	7.7807	8.7629	7.6241	7.6241
IARV	0.9071	0.6052	0.6115	0.6070	0.6335	0.8693	0.6014	0.6014
CR	0.6377	0.8398	0.8385	0.8404	0.8250	0.7988	0.8425	0.8425
ER	0.9193	0.8260	0.8233	0.8233	0.8357	0.8000	0.8261	0.8261
HausdorffDist	2.2811	1.7714	1.7602	1.7509	1.8274	2.0150	1.7510	1.7510
NucleusDist	1.7416	1.3709	1.3731	1.3689	1.3993	1.5711	1.3613	1.3613
BertoluzzaDist	1.4049	1.2365	1.2417	1.2380	1.2485	1.4292	1.2304	1.2304

Table 3.1. Goodness of fit output for interval simple linear regression example (Blood presure values)

Regression Models								
	CM	MinMax	ArithBased	CRM	CCRM	LassoBased	ParamApproach	CCRJM
IR2	0.8615	0.8933	0.8362	0.9016	0.8983	0.8151	0.9104	0.8990
RMSE	0.1312	0.1131	0.1287	0.1022	0.1063	0.1413	0.0977	0.1086
MAPE	1.9668	1.7198	1.9047	1.5866	1.6187	2.1049	1.5490	1.6610
IARV	0.1385	0.1139	0.1254	0.0984	0.1017	0.1849	0.0896	0.1010
IUTheil	1.4420	1.2980	1.3542	1.2109	1.2313	1.6092	1.1661	1.2225
CR	0.6124	0.3495	0.2854	0.2937	0.2691	0.2386	0.3042	0.4569
ER	0.3004	0.2647	0.2994	0.2963	0.2703	0.2301	0.2988	0.3471
HausdorffDist	0.1874	0.1502	0.1722	0.1295	0.1361	0.1758	0.1279	0.1492
NucleusDist	0.1453	0.1249	0.1694	0.1295	0.1361	0.1758	0.1279	0.1492
BertoluzzaDist	0.1271	0.1184	0.1341	0.1147	0.1153	0.1541	0.1128	0.1149

Table 3.2. Goodness of fit output for interval simple linear regression example (Stock exchange values)

3.3.2 Multiple regression example

For a multiple regression, an example of the temperature of some cities during every day of an entire year has been chosen. Córdoba's temperature is going to be predicted by the temperature of Sevilla, Jaen and Ciudad Real knowing the low and high temperatures of each of these cities for every day for the entire year 2014.

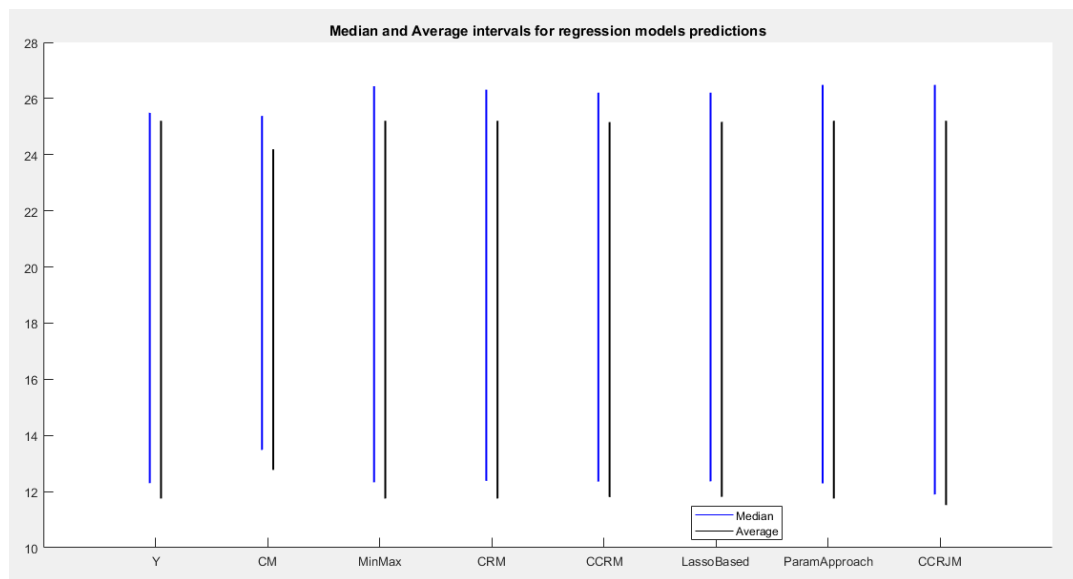


Figure 3.11. Prediction intervals for multiple linear regression example

As seen in Figure 3.11 the output intervals for all methods seem to be close enough to the observed values, once again the numbers of the different metrics have to determine which methods are better.

Looking at Table 3.3 some aspects can be seen:

- CM continues to be the worst for every metric, however in this case the difference with the other methods is not as big as in the simple linear regression.
- By a very small margin, but ParamApproach is still the best method by almost every metric used.

- Lasso has improved its performance with respect to simple linear regression and its on par and close to the top methods by some metrics.
- MinMax is the top model in a couple of metrics and a close second in the rest.
- As explained before, ArithmeticBased is not included as this method its only designed for interval simple regression.

Regression Models							
	CM	MinMax	CRM	CCRM	LassoBased	ParamApproach	CCRJM
IR2	0.8634	0.8863	0.8846	0.8841	0.8841	0.8873	0.8847
RMSE	2.2251	2.0118	2.0285	2.0321	2.0322	2.0027	2.0345
MAPE	14.9205	12.8870	12.5289	12.6073	12.6873	12.7204	13.3372
IARV	0.1366	0.1137	0.1154	0.1159	0.1159	0.1127	0.1153
IUTheil	0.7071	0.6449	0.6499	0.6514	0.6514	0.6420	0.6496
CR	0.7739	0.8898	0.8872	0.8769	0.8770	0.8906	0.8955
ER	0.9165	0.8799	0.8791	0.8836	0.8836	0.8799	0.8707
HausdorffDist	2.7313	2.2253	2.2571	2.2567	2.2561	2.2253	2.2699
NucleusDist	2.1205	1.6994	1.7227	1.7281	1.7282	1.6997	1.7440
BertoluzzaDist	1.5900	1.3788	1.3907	1.3942	1.3942	1.3784	1.4249

Table 3.3. Goodness of fit output for interval multiple linear regression example

Chapter 4

Basic Analytic System for Interval-Valued Data

IN this chapter a walk through the user interface and final tool Basic Analytic System for Interval Valued Data (BASIVD for short) will be made to explain the different screens and functionalities available for the user and how to make them work.

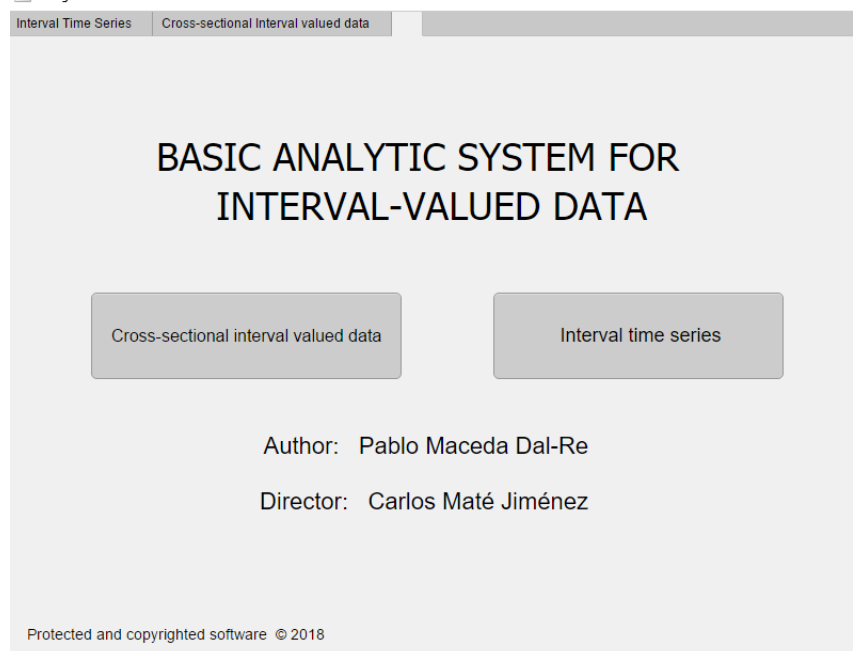


Figure 4.1. Home screen for BATIVD

When the application is run, either from MATLAB or in the standalone application version, the first screen shown will be that of Figure 4.1. It is a simple home page with the title, authors and 2 options for the user to go to: the "Cross-sectional interval valued data" or the "Interval time series" alternative.

4.1 Cross-sectional interval valued data tab

The first thing the user will see for the Cross-sectional interval valued data (CSIVD) is the Visualisation tab Figure 4.2.

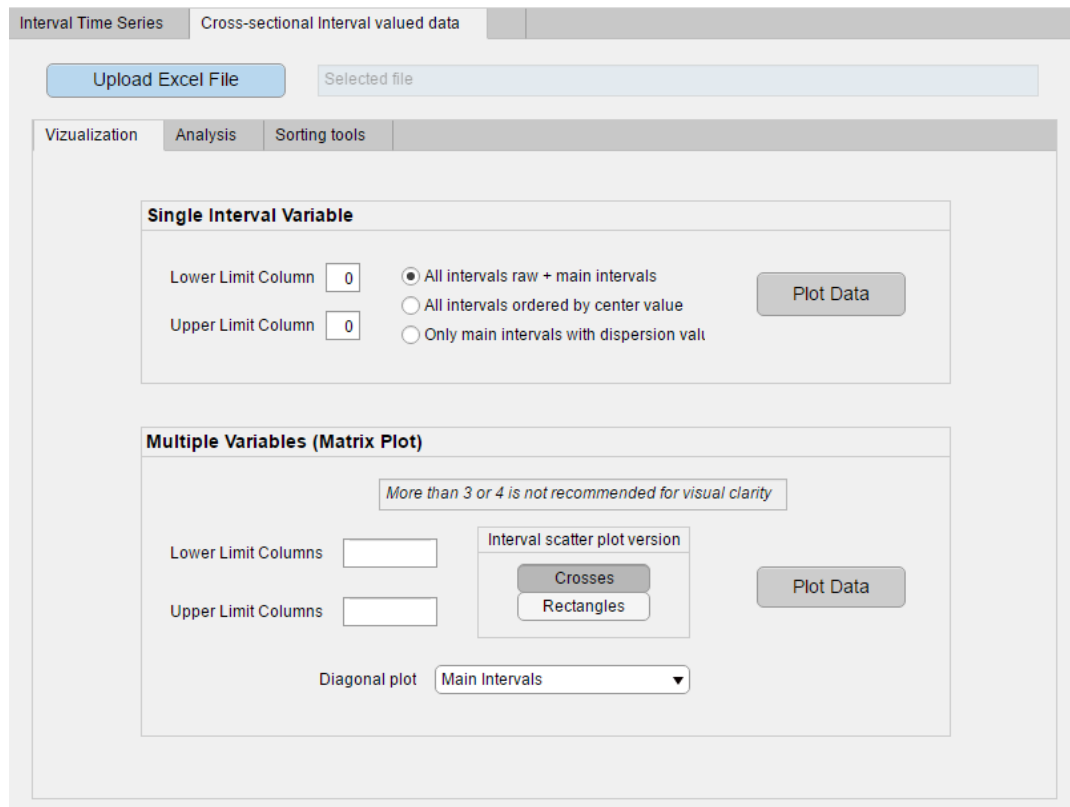


Figure 4.2. Visualisation screen for CSIVD

However, the most crucial part is the "Upload excel file" button on blue on the top. The file uploaded will be common for all CSIVD tabs and, therefore, analysis. It only needs to be uploaded once and it has to

be an Excel (.xlsx) file with the data that will be used for the analysis. When it is correctly uploaded to the tool the blue text box next to the button will change from "Selected File" (Figure 4.2) to show its path and file name like in Figure 4.3.

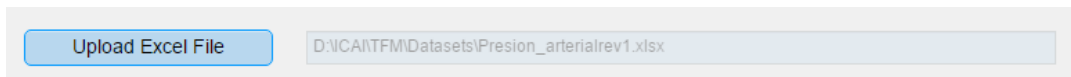


Figure 4.3. Successful file upload

4.1.1 CSIVD Visualisation tab

Once the file is completely uploaded, the first analysis can be ran. The visualisation tab is divided into two different panels:

- **Single Interval Variable:** This section allows to plot the different single variable plots like the ones explained in subsection 2.1.1. The first option plots Figure 2.3. The second option plots Figure 2.4 and the final option plots Figure 2.2. As it only accepts one interval, the input boxes "Lower limit column" and "Upper limit column" only allow a single number each to select one interval.
- **Multiple Variables:** This section plots a Matrix plot like the one on Figure 2.9. It gives the option to the user to select if the scatter plots are the Rectangle version (Figure 2.7) or the Crosses version (Figure 2.6) and gives a choice about the diagonal graphs either to be the MIP, the intervals ordered by centre, the 3D histogram (Figure 2.5) or a scatter plot of centres and radii. As it expects few intervals (not more than 3 or 4 for visual clarity), the input boxes "Lower limit columns" and "Upper limit columns" expect an array of column numbers written within brackets separated by spaces only (e.g. [1 3 5]).

Once the chosen panel has been filled and the plot choice has been chosen, it only requires to press the corresponding button and it will run the analysis.

4.1.2 CSIVD Analysis tab

Figure 4.4. Analysis screen for CSIVD

Without needing to change the file, the user can change to the Analysis tab as shown in Figure 4.4. Here the user will get the chance to run a deep regression analysis and find the best model fit.

This tab shows three panels to be filled:

- **Data Input:** This section allows the user to introduce the desired data columns. For X it can be a single number for single regression or a vector of numbers in between brackets for multiple regression. For Y, as all the methods are univariate, only a single number is allowed for a single variable.

- **Regression Methods:** This section is used to select the models to be run, at least one has to be selected. The only one that won't be run if the regression is multiple is the Arithmetic Based.
- **Fit Measures:** This final section is used to select the fitting measures to be shown in the output table. It also gives the option to select one of them to order the output table's models from best to worst according to that measure. The selected measure to be ordered by has to be selected from the box.

	IR2	RMSE	MAPE	IARV	CR	ER	HausdorffDist	NucleusDist	BertoluzzaDist
CCRJM	0.3986	1.2848	7.6241	0.6014	0.8425	0.8261	1.7510	1.3613	1.2304
ParamApproach	0.3986	1.2848	7.6241	0.6014	0.8425	0.8261	1.7510	1.3613	1.2304
MinMax	0.3948	1.2887	7.6785	0.6052	0.8398	0.8260	1.7714	1.3709	1.2365
CRM	0.3930	1.2917	7.7030	0.6070	0.8404	0.8233	1.7509	1.3689	1.2380
ArithBased	0.3885	1.2972	7.7274	0.6115	0.8385	0.8233	1.7602	1.3731	1.2417
CCRM	0.3665	1.3233	7.7807	0.6335	0.8250	0.8357	1.8274	1.3993	1.2485
LassoBased	0.1307	1.5452	8.7629	0.8693	0.7988	0.8000	2.0150	1.5711	1.4292
CM	0.0929	1.6262	9.5381	0.9071	0.6377	0.9193	2.2811	1.7416	1.4049

Figure 4.5. Output regression methods table

Once all the selections have been done, the button "Run Interval Regression" has to be selected and it will return the fit model lines (Figure 3.8) in case of a single linear regression, the median and average interval from Y and \hat{Y} of all the models (Figure 3.9) and the output table (Figure 4.5) with all the measures selected for all the models ordered by the desired measure (by iARV in this example).

4.1.3 CSIVD Sorting tools tab

The sorting tools tab includes two very useful tools to work with a dataset that might be too big or that the user might want to sort.

- **Sort Dataset by categorical value:** This function allows the user to divide the dataset into different subsets according to one or many categorical variable values. The user must introduce the upper and

The screenshot displays the 'Sorting tools' interface for CSIVD. At the top, there are tabs for 'Interval Time Series' and 'Cross-sectional Interval valued data'. Below these, there is an 'Upload Excel File' button and a 'Selected file' input field. The main area is divided into two sections:

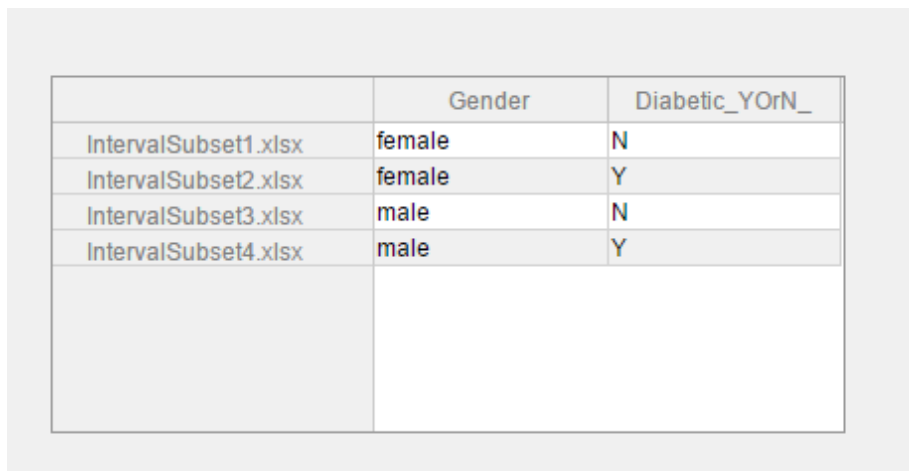
- Sort Dataset by categorical values:** This section contains a descriptive text box stating: "This function will create multiple Excel files as explained by the output table with subsets of the dataset classified by the sorting variables selected by the user. You can use the resulting Excel files for further analysis." Below this are three input fields: 'Lower Limit Column(s)', 'Upper Limit Column(s)', and 'Sorting variable(s) column(s)'. A 'Sort Dataset' button is positioned below these fields.
- Sort Dataset by threshold value:** This section contains a descriptive text box stating: "This function will create two Excel files of interval valued data one with intervals associated to the values over the threshold variable value and other with intervals associated with values under or equal to it." Below this are four input fields: 'Lower Limit Column(s)', 'Upper Limit Column(s)', 'Continuous variable column', and 'Threshold value' (which has the value '0' entered). A 'Sort Dataset' button is positioned below these fields.

Figure 4.6. Sorting tools screen for CSIVD

lower limit columns, as these and only these are going to be the columns that will appear on the new files. And the columns of the categorical sorting variables. Once the button is pressed the tool will create a number of Excel files called "IntervalSubsetX" and return a table like the one on Figure 4.7 to show what values of each sorting variable are represented on each "IntervalSubsetX".

- **Sort Dataset by threshold value:** This function allows the user to take a dataset and divide it according to a threshold value of a continuous sorting variable. The user must introduce the once again the upper and lower limits array of columns and then the column (a single one is allowed in this case) and the threshold value that will make the separation. Once the button is pressed two files will be created, one called "IntervalSubsetOver" with

all the intervals that had a value associated with them larger than the threshold and one called "IntervalSubsetUnder" with all the intervals that had a value associated with them smaller or equal to the threshold. If the files have been created successfully, a message like the one on Figure 4.8 will pop up, if not a caution message will show up.



	Gender	Diabetic_YOrN_
IntervalSubset1.xlsx	female	N
IntervalSubset2.xlsx	female	Y
IntervalSubset3.xlsx	male	N
IntervalSubset4.xlsx	male	Y

Figure 4.7. Sort by categorical variables return index

If the user would like to run different continuous variables filters or mix continuous variable and categorical he will only need to do this functions a number of times. It is important to include on the upper or lower limit column(s) array the variable that wants to sort in the next filter, as the new files will only include the columns selected in those fields. Once the first sorting function in finished, the user just needs to rename the file, as these functions will overwrite a file with the same name, upload that file to the tool and run the new analysis.

4.2 Interval time series tab

To change in between the ITS and CSIVD tabs the user will just need to click among those tabs on the top of the tool. Once it has changed, the

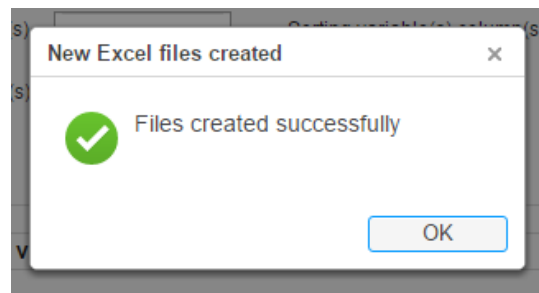


Figure 4.8. Success message from Sort Dataset by threshold value

user will need to upload a new .xlsx file just like explained in section 4.1. Once uploaded correctly it will display the path and file in the same way as in Figure 4.3.

4.2.1 ITS Visualisation tab

Figure 4.9. Visualisation screen for ITS

The visualisation tab is divided into two different panels:

- **Single ITS:** This section allows to plot a single ITS as shown in section 2.2. The user will need to introduce the column where the time variable is included, the upper and lower limit columns and the choice of their joint. If the user chooses None the intervals will be shown just like a vertical line like in Figure 1.4, if they choose centers it will show the vertical lines united by the centers and U & L Limits will show the vertical lines united by the upper and lower limits. The check box is to choose to also plot the graph shown in Figure 2.11 with the increment of center, radius, lower and upper limit from t_{k-1} to t_k .
- **Multiple ITS:** This section allows to plot multiple ITS (more than 4 or 5 is not recommended for visual clarity) as shown in section 2.2. The user will need to introduce the column where the time variable is included, the upper and lower limit columns and the choice of their joint. If the user chooses None the intervals will be shown just like a vertical line like in Figure 2.10, if they choose centers it will show the vertical lines united by the centers and U & L Limits will show the vertical lines united by the upper and lower limits. This function will also return the graph shown in Figure 2.12 to be able to compare different ITS if it not visible in the first plot.

4.2.2 ITS Analysis tab

The Analysis tab for ITS includes the regression panel which is identical and works like the CSIVD one, but it includes another analysis extra.

The **Autocorrelation Function (ACF)** as defined in [20] is a measure of the process' memory at different time lags and is defined by the formula:

Figure 4.10. Analysis screen for ITS

$$ACF = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_i - \bar{y})^2} \quad (4.1)$$

The **Partial Autocorrelation Function (PACF)** gives the partial correlation of a time series with its own lagged values, checking for the values of the series at all shorter lags. It is helpful as the autocorrelation function does not control for other lags.

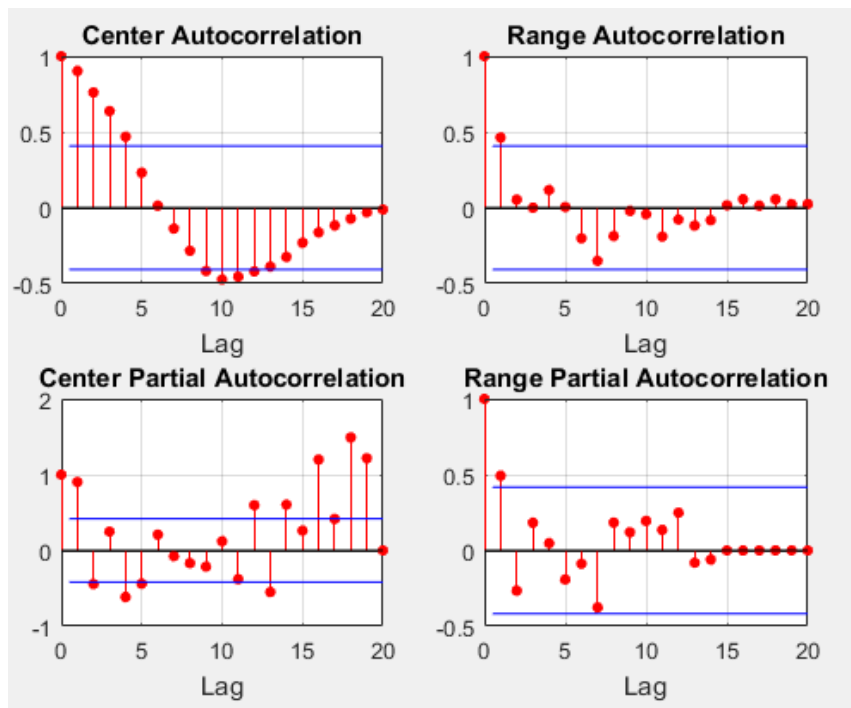


Figure 4.11. ACF and PACF of centres and radii of an ITS

These two analyses are run for the chosen ITS and the resulting plot (Figure 4.11) shows the ACF and PACF of the centres and radii of each ITS, by visually analysing these plots it is possible to determine the lags that are significant (over the blue line).

4.2.3 ITS Cut tool tab

In the Cut tool tab for ITS, the function included allows the user to create a new Excel file with a selected period of the original ITS. The user just needs to introduce the upper and lower limit column values that wants to include in the new file, the time column also to be included and the row from and to it wants to cut the ITS.

If the function works correctly it will return the same success message as the Sort by continuous variables of CSIVD (Figure 4.8).

Upload Excel File D:\ICA\ITFMDatasets\SP500_2008_2009.xlsx

Vizualization Analysis **Cut Tool**

Cut ITS

To select a time period from your ITS, choose the row from and to it spans and the necessary parameters. It will create an excel file called 'CutITS' with resulting section of your ITS to be added for further analysis.

row from Lower Limit Column(s)

Row to Upper Limit Column(s)

Time column

Cut ITS

Figure 4.12. Cut tool screen for ITS

4.3 Tips and installation guide for the tool

4.3.1 Tips

- **File:** The file must be always uploaded and it is the first and only common step for any of the functionalities of the tool. It is a different file for ITS tab and for CSIVD tab, both can be loaded in tool at the same time and work with simultaneously.
- **Column numeric input:** Many functions require column or row values to be able to run. There are two types of input fields: those with a placeholder 0 on it, and those empty. The ones with the place holder 0 are programmed to only accept a single number because that is all that that function needs or uses. The ones empty can accept either a single number (e.g. 5) or an array of

numbers that should ALWAYS be written in brackets with a space in between (e.g [3 14 5]).

- **Multiple sorting filters** If an output file from a sorting function wants to be reused for another filter, it is recommended to rename the file and close the previous one, an error will be given trying to overwrite the file if it is open or the initial one will be deleted if successful. Also the user should try to add to the upper or lower columns not only the limits of the interval variables but any other useful data columns needed, for example another variable to run another filter or analysis in the future.
- **Running time** Be patient with execution times, with large datasets (100+ values) and multiple simultaneous analysis, depending on your computer's processing power it might take a few dozens of seconds to run the desired function, specially the first execution after opening.

4.3.2 Installation and running

There is two ways of installing this tool and running it.

4.3.2.1 MATLAB owners

This is the recommended way for anyone owning a MATLAB version 2016a or later. The tool will be in a MATLAB App Installer (.mlappinstall) format. The user just needs to go to his APPS tab in MATLAB and press "Install App" button as seen in Figure 4.13.

Once the file browser pops up, the user just needs to select the .mlappinstall and it will install in a couple of seconds. After installed, the app will show up in the app directory under "MY APPS" as seen in Figure 4.14 and can be used by just clicking on it.

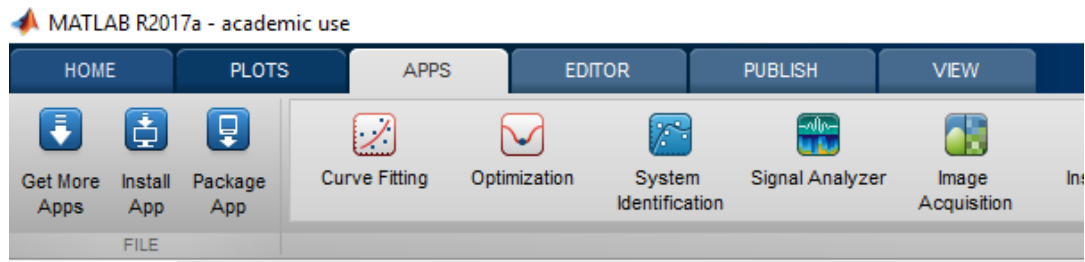


Figure 4.13. MATLAB app installer

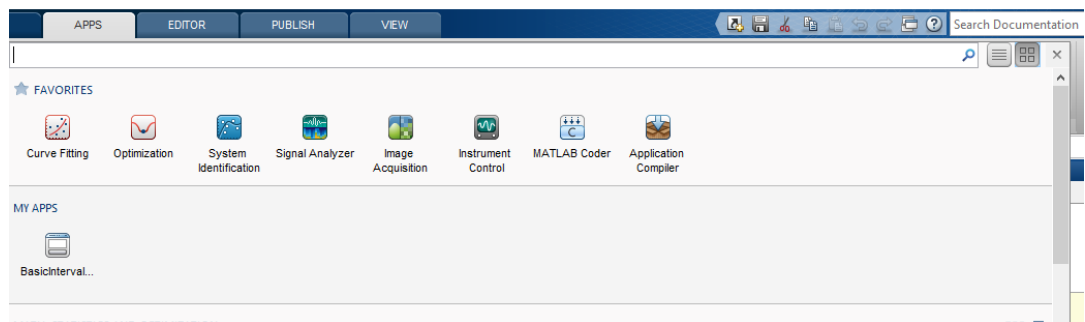


Figure 4.14. MATLAB app installer

4.3.2.2 Non-MATLAB owners

For those that do not own a MATLAB license or it is too old for this tool, there is a standalone application created by the MATLAB App Compiler.

This option will create a much larger .exe file that can be run in any computer, however, its installation will require internet connection to download many functions from Mathworks and much more time, it might take 20 or 30 minutes depending on your hardware.

The good news is that once the tedious installation process is completed, the BASIVD Tool will be added to your computer main menu and can be clicked and used like any other app such as Word, Chrome or Outlook.

For this option the user just needs the .exe file that includes the application. The remaining files needed will either be included in that file or downloaded automatically during the installation from the internet

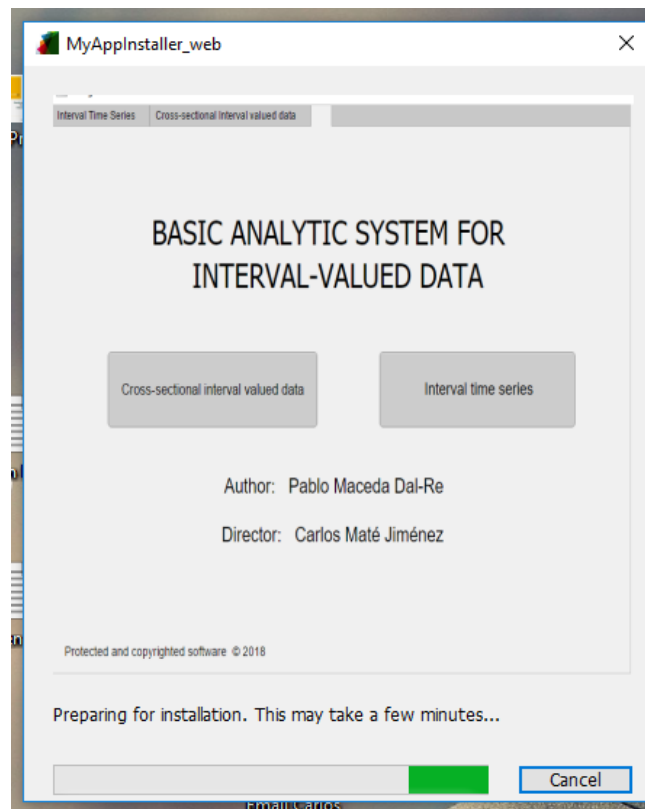


Figure 4.15. Installation as a standalone application

through Mathworks.com. This includes functions programmed by the developer, like many from this project, that MATLAB automatically includes in the .exe file.

This tool is very useful as by only sharing a single .exe file of 15 MB, the app can be shared with anyone willing to use it even if they do not own a MATLAB license.

Chapter 5

Conclusions

IN this chapter a final overview of the project will be made with focus in the main takeaways from it.

5.1 Visualisation

For the visualisation part, this project included new and innovative ways of representing interval-valued data as explained in chapter 2 that were new in the field of interval-valued data.

5.2 Regression

As seen by the examples in section 3.3, the ranking of the different regression methods is not clearly determined.

CM method, as the first one, is very simplistic and its approach in most examples seems to be basic given the interval's nature. However, for intervals with very little radius its centric only approach might be enough and a good approximation as a basic initial approach.

ParamApproach seems to be the best method so far. Its deep approach using both upper and lower limits for the models of upper

and lower limits seems to give a performance boost to its projections. CCRJM performs on a similar note but slightly worse on some cases.

LassoBased seems to give slight worst results for simple linear regression, its parsimony approach works hurts a little. With multiple linear regression however it evens up with the other methods.

The remaining group of methods perform very similarly and are almost interchangeable performance-wise. CCRM might get the edge as it can prevent results that have no valid mathematical explanation.

To sum up, there is no clear preferred method yet for interval regression. ParamApproach seems to be the best performer but it measures barely 1% or 2% better than MinMax method that was developed 15 years earlier. The user has to decide if the extra few percentage points of accuracy are worth the complexity and computational time of the modern methods like CCRJM or ParamApproach. If not, classic methods like CCRM, MinMax or even CM can be good enough for a inial basic analysis.

Bibliography

- [1] **Carlos Maté Jiménez**, *El análisis de intervalos: aplicación a la ingeniería*, June 2012, Anales.
- [2] **Billard, L. and Diday, E.**, *Regression analysis for interval-valued data.*, 2000, Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies. Pages 369–374.
- [3] **Billard, L. and Diday, E.**, *Symbolic regression analysis.*, 2002, Classification, Clustering and Data Analysis, Proceedings of the Eighth Conference of the International Federation of Classification Societies. Pages 281–288.
- [4] **Eufrásio de A. Lima Neto and Francisco de A.T. de Carvalho**, *Centre and Range method for fitting a linear regression model to symbolic interval data*, 2008, Computational Statistics and Data Analysis 52 Pages 1500-1515.
- [5] **Eufrásio de A. Lima Neto and Francisco de A.T. de Carvalho**, *Constrained linear regression models for symbolic interval-valued variables.*, 2010, Computational Statistics and Data Analysis 54, p. 333-347.
- [6] **Beatriz Sinova and Ana Colubi and María Ángeles Gil and Gil González-Rodríguez**, *Interval arithmetic-based simple linear*

- regression between interval data: Discussion and sensitivity analysis on the choice of the metric*, 2012, Information Sciences 199 p. 109–124.
- [7] **Leandro C. Souza and Renata M.C.R. Souza and Getúlio J.A. Amaral and Telmo M. Silva Filho**, *A Parametrized Approach for Linear Regression of Interval Data*, 2017, Knowledge-Based Systems.
- [8] **J. Arroyo and R. Espínola and C. Maté**, *Different approaches to forecast interval time series: a comparison in finance*, 2011, Computational Economics.
- [9] **Laura Morell Merino**, *Forecasts combination system. Application to financial markets using interval time series.*, 2012, Universidad Pontifica Comillas.
- [10] **Javier Redondo Alastrué**, *Interval Time Series Analysis and Forecasting*, 2013, Universidad Pontifica Comillas.
- [11] **Siegfried M. Rump**, *INTLAB - INTERVAL LABORATORY*, 1999, Developements in Reliable Computing.
- [12] **Alvin C. Rencher**, *Methods of Multivariate Analysis*, 1995, Brigham Young University.
- [13] **Eufrasio de A. Lima Neto / Claudio A. V. de Souza Filho / Pedro R. D. Marinho**, *iRegression-package in R*. Published 07-2016 [Documentation](#)
- [14] **Peng Hao, Junpeng Guo**, *Constrained center and range joint model for interval-valued symbolic data regression*, 2017, Computational Statistics and Data Analysis 116 p. 106–138.

- [15] **Paolo Giordani**, *Linear regression analysis for interval-valued data based on the Lasso technique*, 2015, Sapienza University of Rome.
- [16] **Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.**, *A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread*, 2009, Information Sciences 179 p. 3964–3972.
- [17] **Tibshirani, R.**, *Regression shrinkage and selection via the lasso*, 1996, Journal of the Royal Statistical Society - Series B 58, p. 267–288.
- [18] **Ángela Blanco Fernández**, *Análisis estadístico de un nuevo modelo de regresión lineal flexible para intervalos aleatorios*, 2009, Universidad de Oviedo.
- [19] **Carlos Maté Jiménez**, *Combining interval time series forecasts. An agenda for future research*, July 2015, 1st International Symposium on Interval Data Modelling: Theory and Applications.
- [20] **A. Ullah and D. Giles.**, *Handbook of empirical economics and finance*, 2011, CRC Press.

Acronyms

<i>ABSLR</i>	Arithmetic-based simple linear regression
<i>ACF</i>	Autocorrelation Function
<i>BASIVD</i>	Basic Analytic System for Interval Valued Data
<i>CCRJM</i>	Constrained center and range joint method
<i>CCRM</i>	Constrained Centre and Range Method
<i>CM</i>	Centre Method
<i>CR</i>	Coverage Rate
<i>CRM</i>	Centre and Range Method
<i>CSIVD</i>	Cross-sectional interval valued data
<i>ER</i>	Efficiency Rate
<i>iARV</i>	Interval average relative variance
<i>ILR</i>	Interval Linear Regression
<i>iR²</i>	Interval coefficient of determination
<i>ITS</i>	Interval Time Series
<i>iUTheil</i>	Interval Theil's U statistic
<i>LASSO</i>	Least Absolute Shrinkage and Selection Operator
<i>MAPE</i>	Mean absolute percentage error
<i>MIP</i>	Main Intervals Plot
<i>NNLS</i>	Non-negative least-square
<i>PACF</i>	Partial Autocorrelation Function
<i>RIS</i>	Random interval-valued set
<i>RMSE</i>	Root-mean-square error