

SORT

Statistics and Operations Research Transactions

Volume
45

Number 1, January-June 2021



Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 45, Number 1, January-June 2021

ISSN: 1696-2281
eISSN: 2013-8830

Invited article

The radiant diagrams of Florence Nightingale

Michael Friendly and RJ Andrews

Articles

Verifying compliance with ballast water standards: a decision-theoretic approach

Eliardo G. Costa, Carlos Daniel Paulino and Julio M. Singer

Bayesian classification for dating archaeological sites via projectile points

Carmen Armero, Gonzalo García-Donato, Joaquín Jimenez-Puerto, Salvador Pardo-Gordo and Joan Bernabeu

Joint outlier detection and variable selection using discrete optimization

Mahdi Jammal, Stephane Canu and Maher Abdallah

The unilateral spatial autogressive process for the regular lattice two-dimensional spatial discrete data

Azmi Chutoo, Dimitris Karlis, Naushad Mamode Khan and Vandna Jowaheer

Information for authors and subscribers



www.idescat.cat/sort/

Aims

SORT (Statistics and Operations Research Transactions) -formerly *Qüestió-* is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society and the Catalan Statistical Society. *SORT* promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts. *SORT* is an Open Access journal which does not charge publication fees.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, *Índice Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa (Qüestió)*, published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

Editor in Chief

David V. Conesa, *Universitat de València, Dept. d'Estadística i Investigació Operativa*

Executive Editors

Esteve Codina, *Universitat Politècnica de Catalunya, Dept. d'Estadística i Investigació Operativa*

Maria L. Durbán, *Universidad Carlos III de Madrid, Depto. de Estadística y Econometría*

Guadalupe Gómez, *Universitat Politècnica de Catalunya, Dept. d'Estadística i Investigació Operativa*

Montserrat Guillén, *Universitat de Barcelona, Dept. d'Econometria, Estadística i Economia Espanyola (Past Editor in Chief 2007-2014)*

Pere Puig, *Universitat Autònoma de Barcelona, Dept. de Matemàtiques (Past Editor in Chief 2015-2020)*

Enric Ripoll, *Institut d'Estadística de Catalunya*

Production Editor

Michael Greenacre, *Universitat Pompeu Fabra, Dept. d'Economia i Empresa*

Editorial Advisory Committee

Jaume Barceló	<i>Universitat Politècnica de Catalunya, Dept. d'Estadística i Investigació Operativa</i>
Eduard Bonet	<i>ESADE-Universitat Ramon Llull, Dept. de Mètodes Quantitatius</i>
Carles M. Cuadras	<i>Universitat de Barcelona, Dept. d'Estadística (Past Editor in Chief 2003–2006)</i>
Pedro Delicado	<i>Universitat Politècnica de Catalunya, Dept. d'Estadística i Investigació Operativa</i>
Josep Domingo-Ferrer	<i>Universitat Rovira i Virgili, Dept. d'Enginyeria Informàtica i Matemàtiques</i>
Paul Eilers	<i>Erasmus University Medical Center</i>
Laureano F. Escudero	<i>Universidad Miguel Hernández, Centro de Investigación Operativa</i>
Josep Fortiana	<i>Universitat de Barcelona, Dept. d'Estadística</i>
Ubaldo G. Palomares	<i>Universidad Simón Bolívar, Dpto. de Procesos y Sistemas</i>
Jaume García	<i>Universitat Pompeu Fabra, Dept. d'Economia i Empresa</i>
Montserrat Herrador	<i>Instituto Nacional de Estadística</i>
Maria Jolis	<i>Universitat Autònoma de Barcelona, Dept. de Matemàtiques</i>
Pierre Joly	<i>Conseil d'Analyse Economique</i>
Ludovic Lebart	<i>Centre Nationale de la Recherche Scientifique</i>
Richard Lockhart	<i>Simon Fraser University, Dept. of Statistics & Actuarial Science</i>
Geert Molenberghs	<i>Leuven Biostatistics and Statistical Bioinformatics Centre</i>
Josep M. Oller	<i>Universitat de Barcelona, Dept. d'Estadística</i>
Javier Prieto	<i>Universidad Carlos III de Madrid, Dpto. de Estadística y Econometría</i>
C. Radhakrishna Rao	<i>Penn State University, Center for Multivariate Analysis</i>
José María Sarabia	<i>Universidad de Cantabria, Dpto. de Economía</i>
Albert Satorra	<i>Universitat Pompeu Fabra, Dept. d'Economia i Empresa</i>
Albert Sorribas	<i>Universitat de Lleida, Dept. de Ciències Mèdiques Bàsiques</i>
Santiago Thió	<i>Universitat de Girona, Dept. d'Informàtica, Matemàtica Aplicada i Estadística</i>
Vladimir Zaiats	<i>Universitat Autònoma de Barcelona, Dept. d'Economia i d'Història Econòmica</i>

Institut d'Estadística de Catalunya

The mission of the Statistical Institute of Catalonia (Idescat) is to provide high-quality and relevant statistical information, with professional independence, and to coordinate the Statistical System of Catalonia, with the aim of contributing to the decision making, research and improvement to public policies.

Management Committee

President

Xavier Cuadras Morató *Director of the Statistical Institute of Catalonia*

Secretary

Cristina Rovira *Deputy Director General of Production and Coordination*

Editor in Chief

David V. Conesa *Universitat de València, Dept. d'Estadística i Investigació Operativa*

Representatives of the Statistical Institute of Catalonia

Cristina Rovira *Deputy Director General of Production and Coordination*
Josep Maria Martínez *Head of Department of Standards and Quality*
Josep Sort *Deputy Director General of Information and Communication*
Josep Jiménez *Head of the Department of Communication and Dissemination*
Elisabet Aznar *Responsible for the Secretary of SORT*

Representative of the Universitat Politècnica de Catalunya

Guadalupe Gómez *Department of Statistics and Operational Research*

Representative of the Universitat de Barcelona

Jordi Suriñach *Department of Econometrics, Statistics and Spanish Economy*

Representative of the Universitat de Girona

Santiago Thió *Department of Informatics, Applied Mathematics and Statistics*

Representative of the Universitat Autònoma de Barcelona

Xavier Bardina *Department of Mathematics*

Representative of the Universitat Pompeu Fabra

David Rossell *Department of Economics and Business*

Representative of the Universitat de Lleida

Albert Sorribas *Department of Basic Medical Sciences*

Representative of the Universitat Rovira i Virgili

Josep Domingo-Ferrer *Department of Computer Engineering and Maths*

Representative of the Catalan Statistical Society

Núria Pérez *Fight Against AIDS Foundation*

Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
Fax. +34 - 93 557.30.01
E-mail: sort@idescat.cat

Publisher: Institut d'Estadística de Catalunya (Idescat)

© Institut d'Estadística de Catalunya

ISSN 1696-2281

eISSN: 2013-8830

DL B-46.085-1977

Key title: SORT

Numbering: 1 (december 1977)

www.idescat.cat/sort/



ISSN: 1696-2281

eISSN: 2013-8830

SORT 45 (1) January-June (2021)



SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya

Universitat de Barcelona

Universitat de Girona

Universitat Autònoma de Barcelona

Universitat Pompeu Fabra

Universitat de Lleida

Universitat Rovira i Virgili

Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society

Societat Catalana d'Estadística



Generalitat
de Catalunya
**Institut d'Estadística
de Catalunya**

SORT

Volume 45

Number 1

January-June 2021

ISSN: 1696-2281

eISSN: 2013-8830

Invited article

- The radiant diagrams of Florence Nightingale (invited article) 3
Michael Friendly and RJ Andrews

Articles

- Verifying compliance with ballast water standards: a decision-theoretic approach 19
Eliardo G. Costa, Carlos Daniel Paulino and Julio M. Singer
- Bayesian classification for dating archaeological sites via projectile points 33
Carmen Armero, Gonzalo García-Donato, Joaquín Jimenez-Puerto, Salvador Pardo-Gordo and Joan Bernabeu
- Joint outlier detection and variable selection using discrete optimization 47
Mahdi Jammal, Stephane Canu and Maher Abdallah
- The unilateral spatial autogressive process for the regular lattice two-dimensional spatial discrete data 67
Azmi Chutoo, Dimitris Karlis, Naushad Mamode Khan and Vandna Jowaheer

The radiant diagrams of Florence Nightingale

Michael Friendly¹ and RJ Andrews²

¹ Psychology Department, York University. ² Independent Author.

Abstract

This article is a tribute to the contributions of Florence Nightingale to statistics and statistical graphics on her bicentennial. We start with her most famous “rose” diagram and describe how she came to this graphic, designed to influence medical practice in the British army. But this study takes us backward in time to consider where and when the ideas of radial diagrams arose, why they were useful, and why we call these her “radiant diagrams.”

MSC: 62-03, 62-09.

Keywords: Data visualization, polar area diagram, radial diagram, nursing, sanitation.

Introduction

This article is a celebration of Florence Nightingale (FN), on the slightly belated occasion of the 200th anniversary of her birth on May 12, 1820, but in time for the *International Year of Women in Statistics and Data Science: A Tribute to Florence Nightingale*, being promoted by many statistical societies worldwide. In her time, she achieved prominence as a reformer of hygiene in hospitals and medical practice, motivated by her experience in the Crimean War. She became known as the “Lady with the Lamp”¹ and is today considered the mother of modern nursing. Mobile Army Surgical Hospitals (MASH units) are part of her legacy, recounted in the eponymous TV series.

However, it is her pen rather than her lamp we pay tribute to here. Following her time in the Crimea, she launched a campaign to further the cause of army hospital reform and wielded impressively detailed data and radiant diagrams to convince those with influence in the merit of her cause. The lady with the lamp became a “passionate statistician.”²

In the popular appreciation of FN’s statistical work, she is most well-known for the singular *Diagram of the Causes of Mortality in the Army of the East* that appeared in 1859 (Figure 1).

1. This phrase comes from an 1857 poem by H. W. Longfellow: “Lo! In that house of misery / A lady with a lamp I see”.

2. This phrase is attributed to Edward T. Cook’s 1913 biography, *The Life of Florence Nightingale*. Her biography as a statistician is told by Kopf (1916).

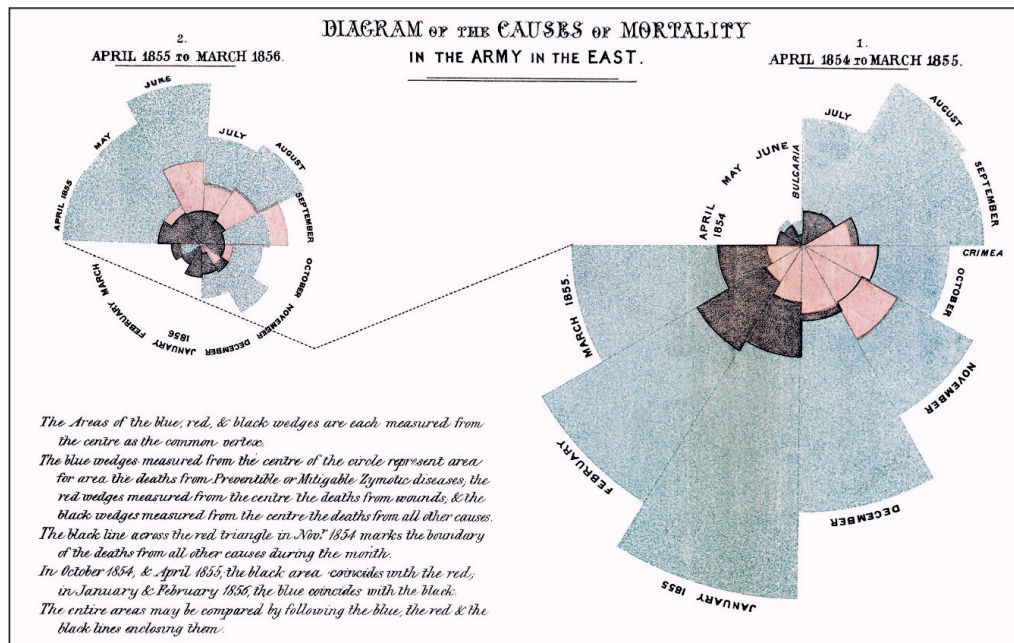


Figure 1: Nightingale's radial diagram of mortality, showing the number of deaths from preventable zymotic diseases (outer, blue wedges), compared with deaths from wounds (pink), and from all other causes (dark gray). Source: Nightingale (1859, p. 19).

The full story of her contributions to visual design and graphic rhetoric is fascinating (Andrews, 2019; Brasseur, 2005) but would take us too far afield from this brief tribute. Rather, we focus on the historical antecedents of her radiant radial diagram, some steps that led her to this, and other diagrams that followed her inspiration.

Life and career

Nightingale was born to a wealthy, landed British family. As a young girl, she exhibited an interest in and flair for mathematics, encouraged by her father, William. One of her mathematics tutors was the renowned James Joseph Sylvester [1814–1897], a contributor to the theory of matrices. Later, she was profoundly influenced by reading Adolphe Quetelet's 1835 *Sur L'Homme et le Developpement de ses Facultés*, in which he outlined his conception of statistical method as applied to the life of man. She also felt a strong religious calling to the service of others, and against her mother's strenuous objections, she decided that nursing would be her vocation.

The Crimean War

The Crimean War was fought by Russia against the forces of France, Britain, and the remnants of the Ottoman Empire. It began in October 1853, over disputed claims of the

rights of Roman Catholics vs. the Eastern Orthodox Church, and lasted until February 1856. Press reports from the war zone soon enraged the British public. These accounts listed high death tolls and descriptions of dying patients crowded on floors of blood-soaked straw, with vermin-infested laundry. In short, the field hospitals were killing British soldiers faster than the enemy. Blame was placed on the government and the military.

The British government knew it had to react. In October 1854, Nightingale appealed to her friend Sidney Herbert, secretary of state for war, to send her and a team of nurses to the Crimea. She soon recognized that most of the deaths occurred, not from battle, but from preventable causes: zymotic diseases (mainly cholera) and insufficient sanitary policy in the hospitals that treated the soldiers.

The Sanitary Commission

Nightingale was more appalled by what she witnessed in the Crimea than what she had read in the newspapers. She developed a system to keep meticulous records of the causes of mortality among the British troops. Her initial attempts to understand these data through tables and charts led to shocking comparisons: deaths in the first seven months of the Crimean campaign amounted to an annual 60% mortality from disease alone. This exceeded that of the Great Plague in London (1665-1666) and that of cholera epidemics in 1848 and 1854. Following her persistent requests to the War Office, a Sanitary Commission was formed around April 1855 to investigate the causes of high mortality of the British Army in the Crimea.

The Royal Commission

After her return to England in July 1856, Nightingale pressed the government (with some support from Queen Victoria) to establish a Royal Commission to examine the causes of mortality in the army. She submitted a report with many tables and concrete proposals for reform, but little was done. How could she turn her insight from experience and data into a powerful call for action?

She met and was befriended by William Farr, the chief statistician of the General Register Office (G.R.O.) established by Parliament to track births and deaths. Farr had become influential in reporting on deaths due to cholera (Farr, 1852), and became an advocate for the careful use of data toward the goal of improving the health of the nation. Farr and Nightingale worked together to access and organize data from the Crimea, systematically analyze it with the help of Farr's team of G.R.O. clerks, and produce persuasive arguments in the form of a series of publications with corresponding text, tables, and diagrams. Farr was an accomplished presenter of statistical "reports." Nightingale elevated their collaborative craft to new heights with her infectious motivation to persuade the British government to adopt sweeping reforms to the entire treatment of soldiers. She said, "The main end of Statistics should not be to inform the Government as

to how many men have died, but to enable immediate steps to be taken to prevent the extension of disease and mortality.” (Nightingale, 1858a, p. 329).

Compared to what?

Nightingale’s most celebrated diagram (Figure 1) was just one of several attempts by her and others to portray the deaths among British soldiers in the Crimea in a way that would capture attention of her readers and provide motivation for a call to action. To understand her graphic design, one key rhetorical question that permeates this work is “compared to what?”³ She broke new ground here in several interesting ways.

Initially, she had just total mortality data, month by month in selected field hospitals of the East. But, how could she make these results most dramatic? For reasons we describe below, she employed what we would now classify as a “radial, polar area chart”. Unlike a pie chart, which uses sectors of varying angle and equal radius to show amounts, FN’s diagram in Figure 1 uses wedges of equal angle (for the months) and varying radius to portray deaths. Nightingale had no particular name for this chart form, but it is common and acceptable to call them “rose diagrams”

Perhaps the most striking feature of her design of this diagram was the separation of the months into two charts, one (on the right) for the period April 1854-March 1855 and the other for April 1855-March 1856. She could have placed the data for all 24 months in one chart, but her design makes a direct comparison of the deaths before the arrival of the Sanitary Commission with those after. Just a pre-attentive, millisecond glance shows the great difference in size (deaths) between the two portions.

A small puzzle is the arrangement of these two pieces. Normally, one would draw the before/after portions left to right, and in each piece, the initial month would be drawn at 12:00 or 3:00. But this made it more difficult to connect the data for March 1855 with that for April, the following month. Her right-to-left design, starting each diagram at 9:00, made it easier to connect these adjacent months with a dotted line.

FN’s earlier attempts

We now consider how Nightingale arrived at the well-known diagram of mortality shown in Figure 1. She had seen a polar diagram in William Farr’s 1852 report on potential causes of mortality due to cholera and was much impressed. In this (Figure 2) he drew circular diagrams showing weekly temperature and cholera deaths in London over the period 1840-1849, as if to establish some link between the two. This kind of chart is sometimes called a *radar chart* today. It uses annular rings with radii proportional to the a given measure; alternatively: a time-series line graph in polar coordinates, where the radial lines serve as axes for the 52 weeks of the year.

3. JW Tukey quote: “The purpose of [data] display is comparison (recognition of phenomena), not numbers...”



Figure 2: Farr's radial diagram of temperature and mortality in London by week for the years 1840-1850. The yearly charts are arranged row-wise from 1840 at the top left. The chart at the bottom right corner shows the average over the years 1840-1849. Outer circles show weekly deaths; inner circles show weekly temperature. Source: Farr (1852), plate IV.

The outer charts show average weekly deaths, relative to the mean over all years, shaded black when they exceeded the average (excess mortality), and yellow otherwise. It was immediately apparent that something horrible had happened with cholera deaths in London in the summer of 1849 (row 3, column 2). But cholera deaths had also spiked in the winter months in 1847 (row 2, column 3).

Farr was searching for easily found associations with cholera mortality here. No direct link to temperature or other factors that he tried (e.g., elevation above the Thames) could be found, until John Snow (1855) argued for a water-born causative agent. But for FN, this radial diagram form seemed exciting and novel; something she could use to make her case.

The Bat-Wing Diagram

Nightingale was impressed enough with Farr's use of a radial diagram to adopt this form for her own data. In her first version (Figure 3), printed privately for the Secretary of War

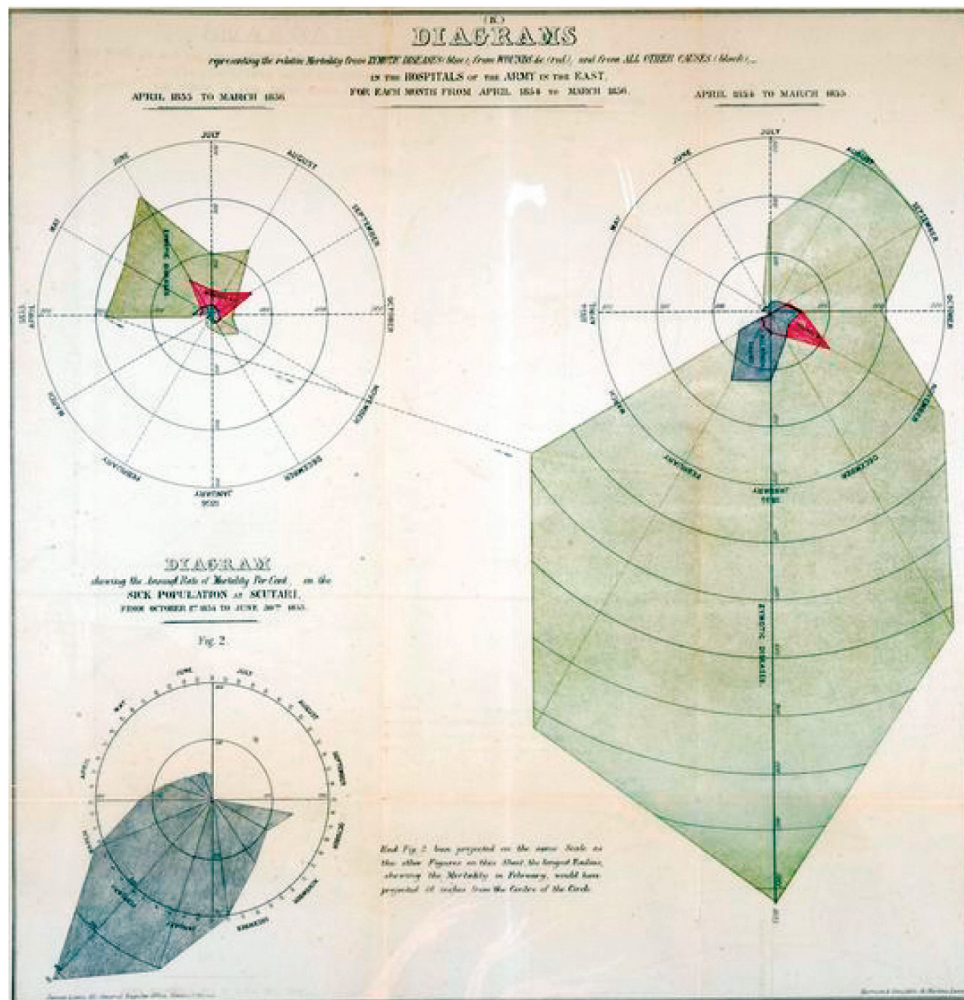


Figure 3: Initial design of Nightingale's diagram, using a linear scale. The two diagrams at the top show relative deaths from zymotic diseases, wounds (red) and other causes (black). The bottom left diagram shows the annual rate of mortality at Scutari from October 1854 to June 1855. Source: Nightingale (1858b, Diagram K, p. 47).

in 1858, she followed Farr's design, which plotted deaths on a linear scale (of deaths per 1000) as distances from the origin, with radial axes corresponding to 100, 200, 300, ...

What she saw here was beyond astounding. The deaths from preventable causes (zymotic diseases) totally swamped those from battle wounds or other causes, and totally dominated the scale. In her Fig. 2 at the bottom left in our Figure 3, she shows the annual rate of mortality in the sick population of Scutari, where the fraction reached 415% in February by her calculations. Here she notes, "Had Fig. 2 been projected on the same scale as the other figures on this sheet, the longest radius, showing the mortality in February, would have projected 40 inches from the centre" (Nightingale, 1858b).

She quickly realized that although the data were correct, this graph was deceptive, because the eye tends to perceive the area rather than length in such displays: doubling the death rate would give a perceived area four times as large. In her subsequent versions, Nightingale plotted deaths in each month as the square roots of distance from the center, so the **area** of each wedge reflected the number of deaths. It is easily seen that deaths from preventable diseases (the outer blue wedges in Figure 1) totally dominate those from battlefield wounds and other causes. This was yet another aspect of her graphical insight that “compared to what” meant that meaningful comparisons had to be on a reasonable scale.

The Manchester Rose

In other earlier versions, Nightingale tried different definitions of “compared to what,” to make her argument salient. Figure 4 is stylistically similar to Figure 1, except that the smaller dotted circles represent “what the mortality would have been for the whole year if the army had been as healthy as men of army age are in Manchester, which is one of the most unhealthy towns in England” (Nightingale, 1859). Her goal here, as in other versions, was to shock the viewer: compared to even Manchester, the army in the East was suffering unfathomable losses.

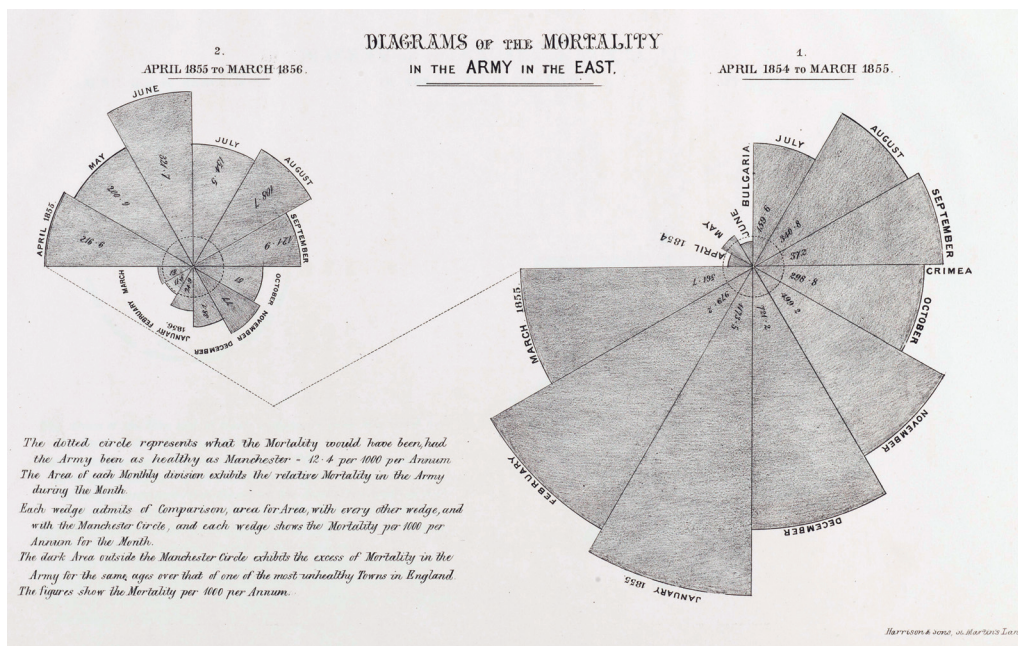


Figure 4: Diagram of the mortality in the British Army in the east during April 1854 to March 1855 (right) and April 1855 to March 1856 (left) in comparison to that of Manchester, represented by the circular dotted line. Source: Nightingale (1859), p. 320.

Radial diagrams before FN

We described earlier how the immediate stimulus for Nightingale's use of radial diagrams developed from what she learned from Farr (Figure 2) and how she discovered that such diagrams were more perceptually accurate when counts of deaths were presented on a square root scale, so that wedges had areas proportional to the count (Figure 4). But while Nightingale is often credited as the inventor of such charts, it is useful to consider earlier origins.

Guerry's Cycles

As we have argued elsewhere (Friendly, 2007, 2007b), the earliest direct precursor of Nightingale's rose diagram appeared in an 1829 publication by André-Michel Guerry. His goal here was to try to determine if relationships among meteorological variation and physiological phenomena could be found by graphical means; but particularly to show how these could be represented as cyclical phenomena, over months of the year, hours of the day, days of the week and so forth.

Weather phenomena included wind direction, temperature, days of thunder, frost, rain, snow, etc. Physiological phenomena were comprised of various causes of admission to hospital. He also included data on weddings, mortality, suicides by month, and hourly data on births and deaths.

Figure 5 shows the portion of his diagram using the radial wedge form to show average trends for some periodic phenomena at different scales; he called these "courbes circulaires," meaning he saw them as curves wrapped around a circle. The top row here shows average wind directions for four quarters of the year, using the conventional compass orientations. He says:

We have represented by these circular areas, and from the observations of 9 years, the number of days that the various winds blow in Paris during a three-month period... According to popular opinion, the south winds prevail especially in summer, northerly winds in winter. We see that the exact opposite is happening.

This establishes his idea that diagrams of cyclical phenomena can reveal consistencies not easily seen in tables. Graphical methods were still on the rise in 1829. To cite an authority, and frame his study in a wider context, he quotes von Humboldt's (1813) memoir on finding lines of isotherms.

The use of graphic means will throw a lot of light on phenomena of the highest interest. If, instead of geographical maps, we had only latitude, longitude and height coordinates, a large number of curious relationships offered by

the configuration and inequality of the continents would have remained forever unknown.

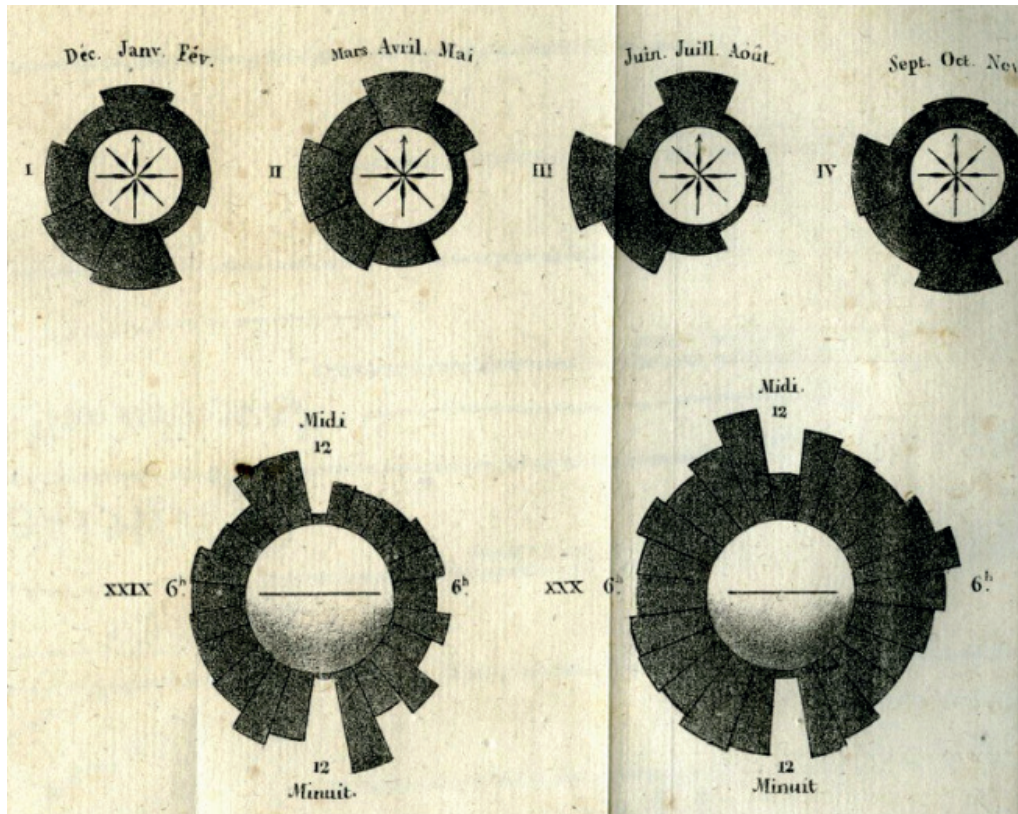


Figure 5: Guerry's radial charts of cyclical phenomena. The top row, charts I-IV. Show the averages of prevailing wind direction by circular area over 9 years according to compass directions. The bottom row, charts XXIX and XXX show, respectively the number of births and deaths over hours of the day. Source: Guerry (1829).

The bottom row in Figure 5 illustrates how he thought that circular diagrams of compass directions could be generalized to other domains. These charts (XXIX and XXX) show variations in births and deaths. He says:

Since the diurnal period represents in some respects the annual period, we have sought if, as with the seasons, there would not be, for some hours, greater ease of births or deaths.

As far as we are aware, this is the first general statement of the graphical principle of radial diagrams for cyclical phenomena, using wedges of constant angle and varying radius.

The French Connection: Guerry → Farr

The link from Farr to Nightingale is clear, but the question arises whether Farr had gotten inspiration for radial diagrams from Guerry. The historical evidence suggests that this is highly likely, though uncertain. What follows is a reasonable account based on our knowledge.

In the early 1800s, following the societal chaos after Napoleon's 1815 defeat, a new idea of "social medicine" or "social epidemiology" began in France (Pinell, 2011). Some leading proponents were Alexandre Parent du Châtelet, Louis-René Villermé and Benoiston de Chateauneuf. In 1829, they launched a new journal, *Les Annales d'Hygiène Publique et de Médecine Légale*, and Guerry published his study in their first volume.

This journal soon became a hub of professional exchange for anyone in the country interested in what was called social hygiene but had a broader scope. It is known that Farr received a bequest in 1828, studied medicine in France and Switzerland, and most likely struck up a friendship with Guerry through the network of the *Annales d'Hygiène Publique*.

Guerry (1833) published his *Essai sur La Statistique Morale de la France*, for which he won the prestigious Montyon Prize upon the recommendation of the Académie Française. In this, Guerry argued that the relations among social and moral variables (literacy, crime rates, suicide, etc.) could be understood using graphs and shaded (choropleth) maps. More importantly he asserted that lawful relations among moral variables could be found, analogous to those of physics. Within a short period of time, this work attracted considerable attention in European statistical circles and Farr was among his admirers.

Guerry's final and most ambitious work was a comparative study of moral variables in England and France which appeared in 1864. Farr is acknowledged for having helped him in obtaining access to court records and other documents in England. In the 30 years between these two works, Guerry displayed his maps and charts in several expositions in Europe. In 1851, he had two exhibitions – an honored public one in the Crystal Palace at the London Exhibition and a second one at the British Association for the Advancement of Science (BAAS) in Bath, England. By October, 1864, Guerry had been made an honorary member of the Statistical Society of London, and was invited again by Farr to attend the BAAS meetings. The *Statistique Morale de l'Angleterre...* (Guerry, 1864) and its splendid plates were put on public display for the nearly 2800 members who attended, and became the subject of a public commentary by W. Heywood, vice-president of the Society.

Farr was not a graphic innovator, but he was tuned-in enough to recognize useful graphical methods and apply them in his work. It is quite likely that his radial diagrams (Figure 2) were inspired by Guerry, and perhaps Léon Lalanne, considered next.

Lalanne's Winds

Another French connection was Léon Lalanne, an engineer of the École Nationale des Ponts et Chaussées (along with Charles-Joseph Minard). Lalanne made several innovations in graphical methods, but the one of interest here is his polar area plot of wind directions (Figure 6). This figure shows the average relative frequency of wind directions recorded at Aigue-Mortes in Occitanie, France over some period of time.

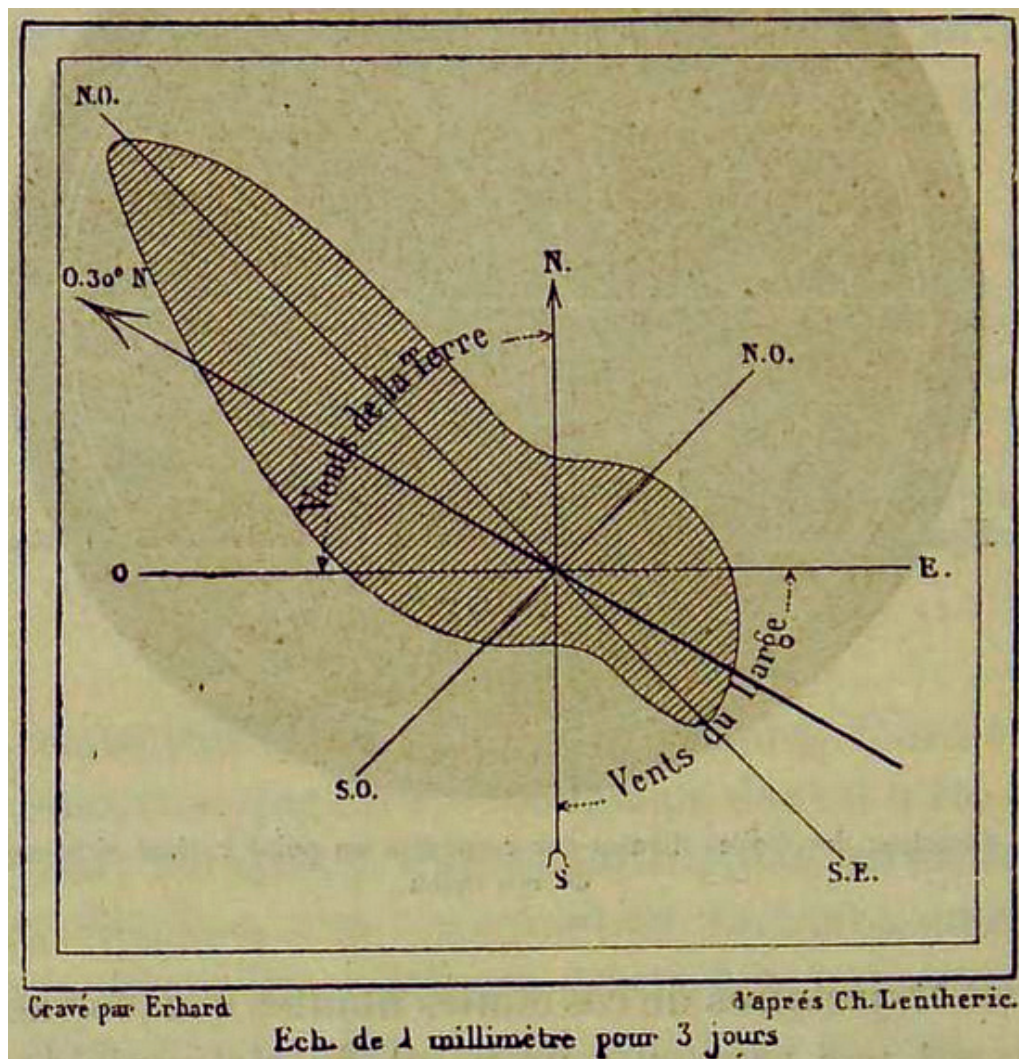


Figure 6: Average prevailing wind directions at Aigue Mortes. The NW quadrant is considered land winds; the SE quadrant, sea winds. An arrow labeled 0.30° N is apparently the average overall wind direction. Source: originally from Lalanne (1843); this rendition from Marey (1885, Fig. 31, p. 68).

He draws attention to compass directions with primary N-S, E-W axes, and adds secondary axes, NW-SE and NE-SW. But the main message is what he has discovered from this diagram: Winds that blow primarily in the NE quadrant he considers land winds (in the direction toward land). Winds that blow toward the SE quadrant are sea winds. An overall average is shown by a large arrow labeled 0.30° N.

What is remarkable here is that the shaded contour is really a smoothed representation of the data and represents another level of sophistication in radial diagrams: a level-curve (iso-contour) of circular data. Earlier, Lalanne had used polar diagrams to display the frequency, duration, and direction of winds over the months of the year near Calcutta, India. The data for individual years were quite variable, but he recognized (following von Humboldt's (1813) isothermal diagrams) a more general principle, that such level curves could be found for other coordinate systems.

The difference consists merely in that the isothermals are applied to points, the existence of which on the surface of the terrestrial globe is real; whilst the curves of the equal duration of the winds in the same place, during the different seasons of the year, are applied to points, whose position on a plane, or a sphere, or a cone, has been determined by pure convention, by a particular choice of co-ordinates to represent two variable elements [p. 514].

Antecedents of polar diagrams

Circular diagrams go back to antiquity, first with spatial directions for an observer of the sun and stars, and later for compass charts, based on a circle of 360° . Fractions of $(0:3) * 1/4$ easily corresponded to N, E, S, W. Intermediate fractions of $1/8$ gave NE, SE, SW, NW. Half-way between these gave NNE, NNW, etc. A navigator could always use direct degrees for a compass heading. Wind directions could be referenced in the same coordinates.

Similarly, the idea of a 24-hour day goes back at least 4000 years, with 12 sections for the night marked by stars that rose and fell, and an equal number of sections for the day. As mechanical clocks developed after the 13th Century, a double 12-hour clock face evolved, synchronized with noon or midday as AM (ante meridiem) and PM (post meridiem). A 12-hour clock face could be divided into $1/4$ fractions (3, 6, 9, 12) or thirds (4, 8, 12).

The origin of pie charts (Spence, 2005) showing parts of a whole is usually traced to Playfair (1801), but there are earlier examples based on clock faces. Among these, the engraving by Nicolas Guérard (undated, but ca. 1700) shown in Figure 7 captures the style and intent in a graphic story illustrated by clock faces.⁴

4. We are grateful to Antoine de Falguerolles for discovering and translating this image.



Figure 7: Clock-face drawing by Nicolas Guérard (1648?-1719) showing a circular (pie chart) representation for compositional data, namely time-budgets. The two clock shields are supposed to represent the paradise for women, and purgatory for men, with the horse in Hell. Source: <https://gallica.bnf.fr/ark:/12148/btv1b8407520q.item>

A hermaphrodite rider (left: woman; right: man) rides a horse, each holding a 24-hour clock representing the way she/he spends a typical day. The content is a totally sexist, deplorable depiction of the daily life of women (left shield) vs. men (right) showing the supposed fractional composition of activities in a day by hours on a clock. Segments are labeled for women (dressing, church, promenade, ...) and for men (different forms of work), but the main visual message is shown by the shaded sectors: 10 hours of repose for women compared with 4 hours for men. Perhaps the title: *Aujourd'hui d'une façon demain de l'autre* (today one way, tomorrow the other) can be read as a call to greater gender equality.

More radiant diagrams

Following Nightingale, the graphical idea of radial diagrams took off, but nowhere in as impressive a form as used by Émile Cheysson, in various volumes of the *Albums de Statistique Graphique*. As Charles-Joseph Minard had demonstrated earlier (Minard, 1858) in his use of pie charts as proportional symbols on a map, Cheysson saw the potential to illustrate time-varying phenomena in a spatial context to make many aspects of the data visually apparent. If we can think of Minard's pie-chart map of consumption of meat in Paris as Playfair 2.0, then surely Cheysson's wedge maps in the *Albums* can be considered Nightingale 2.0.

Paris theaters

Figure 8 is just one example, designed to show the gross receipts in theaters in Paris from 1878 to 1889, but to highlight the influence of the Universal Expositions in 1878 and 1889. Each diagram is positioned on a map of Paris, with a size proportional to the total receipts over all years. This places the diagrams for the theaters in spatial context and allows the eye to easily compare them in size and shape. Clearly, the Opéra was most popular overall, followed by the Opéra-Comique.

Within each diagram, the wedges are area-proportional to receipts in each year, highlighting the exposition years in yellow. The immediate impression is that in both Expo years, more people attended the theaters than in other years.

A rose by any other name

The history of Florence Nightingale and her radial diagrams has many stems and buds. These charts at the time were so novel for her audience that they demanded attention to her essential point: mortality in the army could be decimated by simple medical hygiene measures, just as we all wear masks today to prevent the spread of COVID.

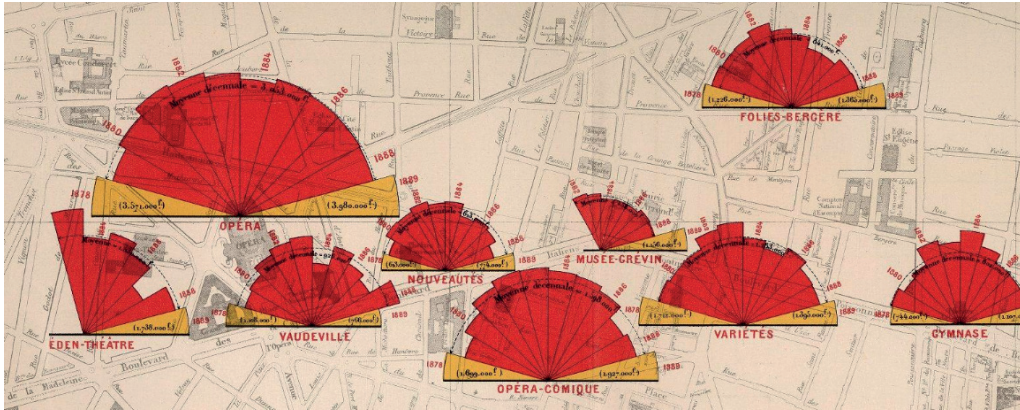


Figure 8: A portion of “Gross receipts of theaters in Paris from 1878 to 1889” (*Recettes brutes des théâtres et spectacles de Paris 1878 à 1889*), highlighting those in the exposition years. Source: *Album de Statistique Graphique*, 1889, Plate 26. [https://www.davidrumsey.com/luna/servlet/detail/RUMSEY 8 1 309502 900 79343:Statistical-Diagram-VI-Exposition](https://www.davidrumsey.com/luna/servlet/detail/RUMSEY%208%201%20309502%2090079343:Statistical-Diagram-VI-Exposition)

But they also seem to demand an equally iconic name. Nomenclature is one stem with multiple buds: “rose”, “coxcomb”, “wedge” diagram are all terms used to refer to these. None of these names have evidence in her writing that she called is such. All of these are somewhat fanciful but attest to a desire to nominate these as a new graphic form.

Here, we announce a new name: **Radiant Diagram** to celebrate FN’s bicentennial and the graphic joy following her footsteps.

In case you were wondering,

- there is indeed a variety of rose called a Nightingale Rose,
- there is also a nightingale bird (*Luscinia megarhynchos*)
- in 1888, Oscar Wilde wrote *The Nightingale and the Rose*, having little to do with our subject, except for its’ lovely alternative title. We might have used this, if it had not already been taken.

References

- Andrews, RJ. (2019). Florence Nightingale is a Design Hero. [Online]. Available: <https://infowetrust.com/project/designhero>.
- Brasseur, L. (2005). Florence Nightingale’s Visual Rhetoric in the Rose Diagrams, *Technical Communication Quarterly*, 14. doi: 10.1207/s15427625tcq1402_3.
- Farr, W. (1852). *Registrar-General, Report on the Mortality of Cholera in England 1848-49*. London: W. Clowes and Sons, for Her Majesty’s Stationery Office.
- Friendly, M. (2007). A.-M. Guerry’s Moral Statistics of France: Challenges for Multivariable Spatial Analysis, *Statistical Science*, 22, 368-399.

- Friendly, M. (2007b). The life and works of André-Michel Guerry (1802-1866). [Online]. Available: <http://www.datavis.ca/papers/GuerryLife.pdf>.
- Guerry, A.-M. (1829). Tableau des Variations météorologique comparées aux phénomènes physiologiques, d'après les observations faites à l'Observatoire royal, et les recherches statistique les plus récentes, *Annales d'Hygiène Publique et de Médecine Légale*, 1, 228-237.
- Guerry, A.-M. (1833). *Essai sur la statistique morale de la France*. Paris: Crochard.
- Guerry, A.-M. (1864). *Statistique morale de l'Angleterre comparée avec la statistique morale de la France, d'après les comptes de l'administration de la justice criminelle en Angleterre et en France, etc.* Paris: J.-B. Baillière et fils.
- Humboldt, A. von (1813): Des lignes isothermes et de la distribution de la chaleur sur le globe. *Memoires de Physique et de Chimie de la Societe d'Arceuil*, vol. 3. Paris
- Kopf, E. W. (1916). Florence Nightingale as Statistician, Publications of the American Statistical Association, 15. [Online]. Available: <http://www.jstor.org/stable/2965763>.
- Lalanne, L. (1843). Appendice sur la representation graphique des tableaux tétéorologiques et des lois naturelles en général. In *Cours Complet de Météorologie*, by L. F. Kaemtzt. Paris: Paulin, 1-35.
- Marey, É.-J. (1885). *La Méthode Graphique dans les Sciences Expérimentales et Principalement en Physiologie et en Médecine*. 2nd Ed. Paris: G. Masson.
- Minard, C.-J. (1858). Carte figurative et approximative des quantités de viande de boucherie envoyées sur pied par les départements et consommées à Paris. [Online] <https://upload.wikimedia.org/wikipedia/commons/1/1c/Minard-carte-viande-1858.png>.
- Nightingale, F. (1858a). *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of The British Army*. London: Harrison and Sons. URL: <https://archive.org/details/b20387118>.
- Nightingale, F. (1858b). *Mortality of the British Army at Home, Home and Abroad, and During the Russian War as Compared with the Civil Population in England*. London: Harrison and Sons.
- Nightingale, F. (1859). *A Contribution to the Sanitary History of the British Army during the Late War with Russia*. London: John W. Parker and Son.
- Pinell, P. (2011). The Genesis of the Medical Field: France, 1795-1870. *Revue française de sociologie*, 5, 117-151. <https://doi.org/10.3917/rfs.525.0117>.
- Playfair, W. (1801). *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe*. London: Wallis.
- Snow, J. (1855). *On the Mode of Communication of Cholera*, 2 ed. London: J. Churchill.
- Spence, I. (2005). No Humble Pie: The Origins and Usage of a Statistical Chart, *Journal of Educational and Behavioral Statistics*, 30, 353-368.

About the Authors

- Michael Friendly is Professor of Psychology at York University, Toronto. He is a Fellow of the American Statistical Association, a developer of data visualization methods and a historian on this topic. His most recent book is *A History of Data Visualization and Graphic Communication*, Harvard University Press, June 2021, co-authored with Howard Wainer.
- RJ Andrews is a data storyteller and editor of *Information Graphic Visionaries*, a book series celebrating data visualization creators. He is author of *Info We Trust, how to inspire the world with data* and guest curator of “Data Visualization and the Modern Imagination”, an exhibit about the history of information graphics at Stanford University.

Verifying compliance with ballast water standards: a decision-theoretic approach

Eliardo G. Costa¹, Carlos Daniel Paulino² and Julio M. Singer³

Abstract

We construct credible intervals to estimate the mean organism (zooplankton and phytoplankton) concentration in ballast water via a decision-theoretic approach. To obtain the required optimal sample size, we use a total cost minimization criterion defined as the sum of the sampling cost and the Bayes risk either under a Poisson or a negative binomial model for organism counts, both with a gamma prior distribution. Such credible intervals may be employed to verify whether the ballast water discharged from a ship is in compliance with international standards. We also conduct a simulation study to evaluate the credible interval lengths associated with the proposed optimal sample sizes.

MSC: 62F15, 62P12.

Keywords: Optimal sample size, Bayes risk, Poisson distribution, negative binomial distribution.

1 Introduction

With the expansion of maritime traffic, ballast water has become the leading dispersing agent of invasive organisms with serious environmental, public health and economic consequences as indicated in Strayer (2010), McCarthy *et al.* (1992) and Marbuah, Gren and McKie (2014). In order to reduce the introduction of invasive species, specially zooplankton and phytoplankton, the international maritime community adopted the Ballast Water Management Convention (BWM Convention) in 2004, that has finally entered into force in 2017. Among other restrictions, the D-2 standard requires that deballasted water should contain no more than 10 viable organisms (referred to simply as organisms in the remainder) with maximum dimension between 10 μm and 50 μm per mL (IMO, 2004).

Given the large amount of ballast water carried by some vessels, it is impractical to analyze the whole water volume and an alternative is to rely on sampling methods that

¹ Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Brazil.

² Departamento de Matemática, IST and CEAUL, FCUL Universidade de Lisboa, Portugal.

³ Departamento de Estatística, Universidade de São Paulo, Brazil.

Received: July 2020

Accepted: February 2021

guarantee some acceptable error rates associated to the decision of whether a given de-ballasting process complies with the D-2 standard. Many authors (First *et al.*, 2013; Carney *et al.*, 2013; Gollasch and David, 2017; Casas-Monroy, Rajakaruna and Bailey, 2020) have addressed this issue, mentioning the quest for “representative” samples, without a consensus on a clear definition and examining samples obtained from a limited number of ship trips.

Very few articles deal with a more structured approach, in which a required sample size is computed to meet some maximum acceptable sampling error (Basurko and Mesbahi, 2011; Miller *et al.*, 2011; Frazier *et al.*, 2013). Costa, Lopes and Singer (2015, 2016), on the other hand, define “representative samples” as those that can be used to estimate the organism concentration in the ballast water tank with a pre-specified precision and use a frequentist approach to compute the optimal sample size with this characteristic. Costa, Paulino and Singer (2021) adopted a Bayesian approach to compute sample sizes required for estimating organism concentration obtained via two optimality criteria: the average coverage and the average length of credible intervals.

As many different tools or methods (*e.g.*, Niskin or Van Dorn bottles, plankton nets, pumps, or the in-line method) may be employed to collect samples from ballast water (Casas-Monroy *et al.*, 2020), it seems reasonable to include costs in the optimal sample size determination. With this in mind, we propose a Bayesian decision approach based on a criterion which minimizes the sum of the sampling method cost and the Bayes risk. An advantage of this approach is that the cost of collecting the sample is explicitly taken into account.

The proposed approach depends on an *ad hoc* loss function defined to accommodate the implications of using a credible interval for the organism concentration λ to decide for compliance or not with the D-2 standard. In a different setup, Etzioni and Kadane (1993) use a similar criterion with quadratic and logarithmic loss functions under a normal model. Sahu and Smith (2006) consider a loss function for the hypothesis testing problem of the parameter of a normal model. Islam (2011) and Islam and Pettit (2012, 2014) consider quadratic, linex and bounded linex loss functions for point estimation of the mean and the variance of a normal model with normal prior distributions, and also exponential and Poisson models both with a gamma prior distribution for point estimation of their respective parameters. Following a similar or the same approach, we may cite Pham-Gia and Turkkan (1992), Bernardo (1997), Lindley (1997), Parmigiani and Inoue (2009), De Santis and Gubbiotti (2017), among others.

Consider a sample $\mathbf{x}_n = (x_1, \dots, x_n)$ consisting of the counts of organisms in n aliquots (sub-samples) with a given volume w collected from a ballast water tank and a specified loss function L . The objective is to obtain the optimal sample size n_o that minimizes a total cost function consisting of the sum of a risk function r and a sampling cost $C(n)$. Once the required optimal sample size has been determined, the corresponding aliquots with volume w are collected (possibly on board or during the deballasting process), the organisms in these aliquots are counted and a credible interval with lower $a(\mathbf{x}_{n_o})$ and upper $b(\mathbf{x}_{n_o})$ limits for the mean organism concentration λ is computed. Considering

that the D-2 standard requires $\lambda < 10$ for compliance, the ship is declared not compliant if $a(\mathbf{x}_{n_0}) \geq 10$ or compliant, if $b(\mathbf{x}_{n_0}) < 10$. Otherwise, if $a(\mathbf{x}_{n_0}) < 10 < b(\mathbf{x}_{n_0})$, more data are needed to make a decision.

In Section 2, we describe two Bayesian models required to compute the credible intervals. The first is appropriate for situations where the organisms are homogeneously distributed in the ballast water tank and the second may be needed for heterogeneous distributions. Sample size determination is presented in Section 3 in terms of a convenient loss function in a decision-theoretic approach. Additionally, we conduct a simulation study to evaluate the lengths of the credible intervals obtained for different combinations of the parameters governing the models and different sampling costs. We conclude, in Section 4, with a discussion of the results and of the difficulties associated to the establishment of the cost components.

2 Bayesian models

2.1 Poisson model with a gamma prior distribution

Let X be the number of organisms in an aliquot of volume w collected from a ballast tank with mean organism concentration λ . The expected number of organisms in this aliquot is $w\lambda$, *i.e.*, $\mathbb{E}[X|\lambda] = w\lambda$. Suppose that, given λ , X follows a Poisson distribution with mean $w\lambda$; this essentially corresponds to the assumption that the organisms are homogeneously distributed in the ballast tank. A possible and first natural choice for a prior distribution is the conjugate gamma distribution for which the density function is

$$h(\lambda) \propto \lambda^{\theta_0-1} \exp(-\theta_0\lambda/\lambda_0),$$

where λ_0 and θ_0 are positive and known fixed constants, respectively interpreted as the prior mean and as a quantity inversely proportional to the prior variance. Thus, the larger (smaller) is θ_0 , the smaller (larger) is the prior uncertainty about λ .

Considering a random sample of size n of $X|\lambda$ and a gamma prior distribution for λ , we may write the model hierarchically as follows

$$X_i|\lambda \stackrel{\text{iid}}{\sim} \text{Poisson}(w\lambda), \quad i = 1, 2, \dots, n; \quad (1)$$

$$\lambda \sim \text{Gamma}(\theta_0, \theta_0/\lambda_0). \quad (2)$$

In this context, the posterior distribution of λ is also a gamma distribution with parameters $\theta_0 + s_n$ and $nw + \theta_0/\lambda_0$, where $s_n = \sum_{i=1}^n x_i$, *i.e.*, $\lambda|\mathbf{x}_n \sim \text{Gamma}(\theta_0 + s_n, nw + \theta_0/\lambda_0)$. Details are presented in the Supplementary Material.

2.2 Negative binomial model with a gamma prior distribution

Suppose that the organism concentration in the i -th aliquot is ℓ_i and the corresponding number of organisms is X_i , $i = 1, 2, \dots, n$. The expected number of organisms in the i -th aliquot is $\mathbb{E}[X_i|\ell_i] = w\ell_i$. For $i = 1, 2, \dots, n$, suppose that, given ℓ_i , X_i follows a Poisson distribution with mean $w\ell_i$ and that given a mean concentration λ in the tank, $\ell_i \sim \text{Gamma}(\phi, \phi/\lambda)$, so that $\mathbb{E}[\ell_i|\lambda] = \lambda$ and $\text{Var}[\ell_i|\lambda] = \lambda^2/\phi$. Thus, given λ and ϕ , X_i follows a negative binomial distribution with $\mathbb{E}[X_i|\lambda, \phi] = w\lambda$ and $\text{Var}[X_i|\lambda, \phi] = w\lambda + (w\lambda)^2/\phi$, where ϕ is a shape (or agglomeration) parameter assumed known (see Amaral Turkman, Paulino and Müller, 2019, Appendix A on the Poisson-gamma mixture). We use the notation $X_i|\lambda, \phi \sim \text{NB}(w\lambda, \phi)$ and again assume a gamma prior distribution for λ .

Considering a random sample of size n from $X|(\lambda, \phi)$ and a gamma prior distribution for λ , we may write the model hierarchically as

$$X_i|\lambda, \phi \stackrel{\text{iid}}{\sim} \text{NB}(w\lambda, \phi), \quad i = 1, 2, \dots, n; \quad (3)$$

$$\lambda \sim \text{Gamma}(\theta_0, \theta_0/\lambda_0). \quad (4)$$

In this context, the posterior distribution of λ is not a known distribution and the computing of its summaries is analytically intractable. Thus, we rely on Markov chain Monte Carlo (MCMC) methods to generate random samples from the distribution of interest. In our case, we use the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) based on a random walk to generate random samples from the posterior distribution of λ . With these samples we may compute related inference summaries. Details are presented in the Supplementary Material.

3 Sample size determination

An approach to the problem of sample size determination and credible interval estimation is to consider it as a decision problem (Lindley, 1997; Parmigiani and Inoue, 2009; Islam and Pettit, 2014). For this purpose, given that λ is the parameter of interest, it is necessary to specify a loss function $L(\lambda, d_n)$ based on a sample $\mathbf{X}_n = (X_1, X_2, \dots, X_n)$ and a decision function $d_n \equiv d_n(\mathbf{X}_n)$. For a given n , the action $d_n(\mathbf{x}_n)$ consists of the specification of two quantities, the lower [say, $a(\mathbf{x}_n)$] and the upper [say, $b(\mathbf{x}_n)$] limits of a credible interval for λ .

Letting $f(\mathbf{x}_n|\lambda)$ be the sampling distribution for \mathbf{X}_n and h be a prior distribution for the unknown parameter λ , the Bayes risk is (see Parmigiani and Inoue, 2009)

$$r(h, d_n) := \int_{\Lambda} \int_{\mathcal{X}^n} L(\lambda, d_n) f(\mathbf{x}_n|\lambda) h(\lambda) d\mathbf{x}_n d\lambda, \quad (5)$$

where Λ is the parameter space and \mathcal{X}^n is the sample space. The Bayes risk $r(h, d_n)$ may be viewed as the mean of the sampling expected loss expressed as a function of

the parameter of interest weighted by the prior distribution; this summarizes the sampling expected loss over all possible values of the parameter of interest (here, the mean concentration λ).

The decision d_n^* that minimizes $r(h, d_n)$ among all possible decisions d_n is called a Bayes rule. Note that if the order of the integration may be inverted, we have

$$\begin{aligned} r(h, d_n) &= \int_{\mathcal{X}^n} \left[\int_{\Lambda} L(\lambda, d_n) h(\lambda | \mathbf{x}_n) d\lambda \right] f(\mathbf{x}_n) d\mathbf{x}_n \\ &= \int_{\mathcal{X}^n} \mathbb{E} [L(\lambda, d_n) | \mathbf{x}_n] f(\mathbf{x}_n) d\mathbf{x}_n, \end{aligned} \quad (6)$$

where $f(\mathbf{x}_n)$ is the marginal distribution of the data, so that the decision d_n^* that minimizes $r(h, d_n)$ is the same that minimizes the posterior expected value of the loss function, namely $\mathbb{E} [L(\lambda, d_n) | \mathbf{x}_n]$, for each \mathbf{x}_n . Given the specified action (the determination of the lower and upper limits of a credible interval for λ in our case), one must define a criterion to obtain an optimal sample size taking both the Bayes risk and the sampling cost into account. With this purpose, we minimize the total cost function $TC(n)$, customarily expressed by

$$TC(n) = r(h, d_n^*) + C(n),$$

where the function $C(n)$ needs to be specified. Here, we take $C(n) = cn$, with c being the cost of sampling an aliquot.

The additive structure of $TC(n)$ in terms of the cost of an action regarding the magnitude of λ and of the sample collection cost presupposes that they are measurable or scalable in some common unit (see Raiffa and Schlaifer, 1961, for example). In fact, we can view $C(n)$ as the relative cost of sampling expressed in terms of the cost associated to the Bayes risk.

Often it is not possible to compute $r(h, d_n^*)$ analytically. In such cases, we may use Monte Carlo simulations as an alternative. Since simulation methods are used, the estimates of $TC(n)$, denoted by $tc(n)$, may show a variation around its true value. We may reduce this variation by: (i) taking the number of Monte Carlo replicates as large as possible and/or, (ii) fitting a curve by least squares or some other method to a set of points $(n, tc(n))$. Müller and Parmigiani (1995) propose to fit the following curve to the estimates of $TC(n)$,

$$tc(n) = \frac{E}{(1 + Hn)^G} + cn,$$

where E , H and G are parameters to be estimated. The numerical methods required to estimate these parameters sometimes do not reach convergence depending on the initial values adopted to implement the corresponding algorithms. In order to simplify the fitting procedure and observing that the parameters H and G play similar roles and essentially represent the decreasing rate of the Bayes risk, we propose to fit the function

$$tc(n) = \frac{E}{(1+n)^G} + cn,$$

that may be linearized as

$$\log[tc(n) - cn] = \log E - G \log(1+n), \quad (7)$$

where the term $-\log(1+n)$ may be interpreted as an explanatory variable and $\log[tc(n) - cn]$, as a dependent variable in a linear regression model. Assuming that an error is added, the estimates of E and G may be computed by least squares. Then, the optimal sample size n_o is the largest integer closest to

$$\left(\frac{\hat{E}\hat{G}}{c} \right)^{1/(\hat{G}+1)} - 1, \quad (8)$$

where \hat{E} and \hat{G} are, respectively, the least squares estimates of E and G .

We adopt the loss function

$$L(\lambda, d_n) = \gamma\tau + (\lambda - m)^2/\tau,$$

where $\gamma > 0$ is a fixed constant, $\tau = (b - a)/2$ is the half-length and $m = (a + b)/2$ is the center of the credible interval (see Rice, Lumley and Szpiro, 2008). The first term involves the half-width of the interval which we may interpret as its precision. The second term, namely, the square of the distance between the parameter of interest (λ) and the center of the interval divided by the half-width to maintain the same measurement unit of the first term, may be interpreted as the bias divided by the precision. If the precision increases (τ decreases) the second term of the loss function increases. The weights attributed to each term are γ and 1, respectively. If $\gamma < 1$, we attribute the largest weight to the second term, prioritizing lower bias over precision; if $\gamma > 1$, the situation is reversed and if $\gamma = 1$, the two terms have the same weight.

For this loss function, the Bayes rule corresponds to the quantities which define the interval $[a^*(\mathbf{x}_n), b^*(\mathbf{x}_n)] = [m^* - \text{SV}_\gamma, m^* + \text{SV}_\gamma]$, where $m^* = \mathbb{E}[\lambda|\mathbf{x}_n]$ and $\text{SV}_\gamma = \gamma^{-1/2}(\text{Var}[\lambda|\mathbf{x}_n])^{1/2}$. For more details see Parmigiani and Inoue (2009), Rice *et al.* (2008) or Schervish (1995).

In a practical situation, once the required optimal sample size n_o has been determined along with the corresponding organism counts \mathbf{x}_{n_o} , the credible interval limits $a^*(\mathbf{x}_{n_o})$ and $b^*(\mathbf{x}_{n_o})$ are obtained via $\mathbb{E}[\lambda|\mathbf{x}_{n_o}]$ and $\text{Var}[\lambda|\mathbf{x}_{n_o}]$, expressed in terms of the models described in the preceding section.

An algorithm to obtain the optimal sample size satisfying the total cost minimization criterion for the adopted loss function is outlined in the Supplementary Material and the corresponding R code is available in Costa, Paulino and Singer (2020). In Tables 1-2 we present optimal sample sizes computed for different values of the parameters defining the prior distributions for both models considered in Section 2. We set λ_0 and obtain the

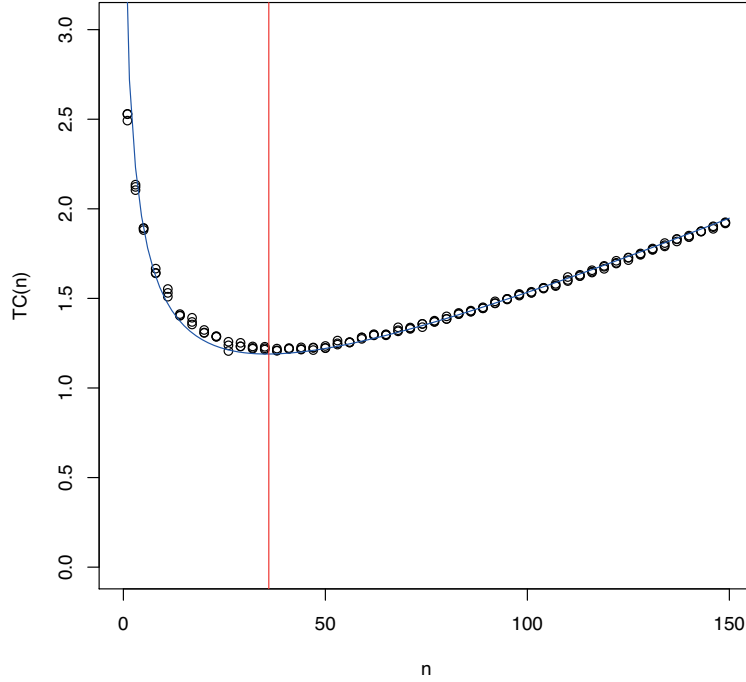


Figure 1: Estimated total cost as a function of n for the negative binomial/gamma model with $\gamma = 1/2$, $\phi = 22$, $w = 1$, $c = 0.01$, $\lambda_0 = 10$ and prior variance equal to 4; the vertical line indicates the optimal sample size $n_o = 36$.

value of θ_0 such that the prior variance is a constant, say, σ^2 , i.e., $\theta_0 = (\lambda_0/\sigma)^2$. See Figure S2 in the Supplementary Material. The values considered for ϕ were chosen to cover the range of estimates obtained from real data reported in Casas-Monroy *et al.* (2020). In Figure 1 we depict a curve fitted to the estimated total cost as a function of n for the negative binomial/gamma model with $\gamma = 1/2$, $\phi = 22$, $w = 1$, $c = 0.01$, $\lambda_0 = 10$ and prior variance equal to 4. The vertical line indicates the optimal sample size $n_o = 36$.

We also carried out a simulation study to evaluate the lengths of the credible intervals and the respective Bayesian coverage probability computed from samples obtained with the proposed optimal sample sizes. For such purposes, we considered the optimal sample sizes obtained via either the Poisson/gamma or the negative binomial/gamma model for combinations of different values of c , θ_0 (and ϕ in the negative binomial/gamma model). For each scenario, we drew 1000 samples \mathbf{x}_{n_o} with the optimal sample size n_o , obtained the limits $a^*(\mathbf{x}_{n_o})$ and $b^*(\mathbf{x}_{n_o})$ of the corresponding credible intervals, computed the mean of their lengths and the mean of the Bayesian coverage probabilities (see Supplementary Material for more details). The results for the average lengths are displayed (within parentheses) in Tables 1-2. The average acceptance rates for the Metropolis-Hastings algorithm used in the negative binomial/gamma model ranged between 31% and 71%. The results for the Bayesian coverage probability are discussed in Section 4.

Table 1: Optimal sample sizes n_o and estimated mean posterior credible interval lengths (within parentheses) under the Poisson/gamma model (1)-(2) with $w = 1$ and $\lambda_0 = 10$.

γ	Aliquot cost (c)	Prior variance		
		1	2	4
1/2	0.001	145 (0.72)	157 (0.70)	164 (0.69)
	0.010	28 (1.45)	32 (1.50)	34 (1.48)
1	0.001	184 (0.45)	198 (0.44)	207 (0.44)
	0.010	35 (0.94)	40 (0.94)	43 (0.94)
2	0.001	237 (0.28)	252 (0.28)	263 (0.27)
	0.010	46 (0.60)	51 (0.60)	55 (0.59)

A simple algorithm with the steps required for the determination of n_o and for the decision with respect to D-2 standard follows.

- Step 1.** Set the values of λ_0 and θ_0 (prior distribution), ϕ (only for negative binomial model), w (aliquot volume), c (aliquot cost) and γ (loss function).
- Step 2.** Obtain the corresponding optimal sample size n_o using the algorithm provided in the Supplementary Material with the parameter values defined in Step 1.
- Step 3.** Sample n_o aliquots of water from the ballast tank of the ship and count the number of organisms in each aliquot. We denote these n_o organism counts as $\mathbf{x}_{n_o} = (x_1, \dots, x_{n_o})$.
- Step 4.** With the organism counts \mathbf{x}_{n_o} and γ compute the credible interval limits $a^*(\mathbf{x}_{n_o})$ and $b^*(\mathbf{x}_{n_o})$ via $\mathbb{E}[\lambda|\mathbf{x}_{n_o}]$ and $\text{Var}[\lambda|\mathbf{x}_{n_o}]$. If there is no closed form for these moments of the posterior distribution, compute estimates for these quantities simulating values from the posterior distribution (using MCMC or another simulation-based method) and taking the respective sample moments.
- Step 5.** Use the credible interval limits to decide for compliance with the D-2 standard as follows: declare compliance if $b^*(\mathbf{x}_{n_o}) < 10$, or non-compliance if $a^*(\mathbf{x}_{n_o}) \geq 10$. Otherwise, if $a^*(\mathbf{x}_{n_o}) < 10 < b^*(\mathbf{x}_{n_o})$, more data are required to make a decision.

4 Discussion

We propose a decision-theoretic approach to obtain an optimal number of aliquots required to estimate the organism concentration in ballast water and indicate how the results may be employed to verify compliance with the D-2 standard.

The results in Table 1 obtained under the Poisson/gamma model indicate that the optimal sample size n_o increases as the prior uncertainty (variance) about λ increases, but the average interval length remains the same. For the negative binomial/gamma model we observe a similar behavior (see Table 2).

Table 2: Optimal sample sizes n_o and estimated mean posterior credible interval lengths (within parentheses) under the negative binomial/gamma model (3)-(4) with $w = 1$ and $\lambda_0 = 10$.

γ	Aliquot cost (c)	ϕ	Prior variance		
			1	2	4
1/2	0.001	1	276 (1.51)	345 (1.48)	347 (1.52)
		4	220 (1.05)	246 (1.03)	259 (1.03)
		8	162 (0.99)	185 (0.96)	206 (0.91)
		13	162 (0.89)	180 (0.87)	192 (0.85)
		22	156 (0.83)	172 (0.80)	182 (0.79)
	0.010	1	-	29 (3.25)	60 (3.18)
		4	21 (2.23)	37 (2.26)	42 (2.35)
		8	26 (1.93)	32 (2.05)	40 (1.99)
		13	25 (1.82)	33 (1.84)	38 (1.82)
		22	23 (1.75)	31 (1.74)	36 (1.71)
1	0.001	1	365 (0.96)	436 (0.95)	458 (0.96)
		4	232 (0.73)	267 (0.70)	292 (0.68)
		8	217 (0.61)	243 (0.59)	260 (0.58)
		13	208 (0.56)	229 (0.55)	244 (0.53)
		22	200 (0.52)	219 (0.51)	231 (0.50)
	0.010	1	-	48 (2.07)	78 (2.04)
		4	34 (1.42)	47 (1.47)	56 (1.47)
		8	36 (1.24)	45 (1.26)	51 (1.26)
		13	35 (1.16)	43 (1.17)	49 (1.15)
		22	35 (1.08)	42 (1.09)	47 (1.07)
2	0.001	1	478 (0.61)	510 (0.63)	678 (0.56)
		4	301 (0.46)	344 (0.44)	373 (0.43)
		8	281 (0.39)	310 (0.37)	331 (0.36)
		13	268 (0.35)	293 (0.34)	309 (0.34)
		22	257 (0.33)	279 (0.32)	292 (0.31)
	0.010	1	-	75 (1.31)	109 (1.27)
		4	61 (0.85)	70 (0.89)	70 (0.95)
		8	51 (0.78)	56 (0.82)	65 (0.80)
		13	51 (0.72)	54 (0.75)	62 (0.73)
		22	49 (0.68)	53 (0.70)	59 (0.68)

For both models and for all n_o in Tables 1-2 the average Bayesian coverage probabilities obtained in the simulation study were approximately 0.84, 0.68 and 0.52 for $\gamma = 1/2, 1$ and 2, respectively. These values are similar to the probabilities that a standard normal variable lies in the intervals $(-\sqrt{2}, \sqrt{2})$, $(-1, 1)$ and $(-1/\sqrt{2}, 1/\sqrt{2})$, respectively, and are consistent with the asymptotic normality of the corresponding posterior distributions. See Ferguson (1996, pg. 140), for example. However, we observe that this approximation also occurs for $n_o \approx 30$. To explain this, first, note that as $\theta_0 \rightarrow \infty$ the gamma distribution approaches a normal distribution (McCullagh and Nelder, 1989, pg. 287). For the Poisson/gamma model (1)-(2) the respective posterior distribution is also gamma with shape parameter $\theta_0 + s_n$, and the cases for which $n_o \approx 30$ are those

where the prior variance is equal to 1 or 2 and correspond to θ_0 equal to 100 or 50, respectively. We may consider these values of θ_0 large enough to guarantee a reasonable approximation by the normal distribution.

The posterior distribution for the negative binomial/gamma model (3)-(4) is neither a gamma distribution nor a known distribution. To verify whether the normal approximation also holds in this case, we considered the smallest sample size in Table 2, namely $n_0 = 21$ and generated 100 samples of size 100 from the posterior distribution of λ . We applied the Shapiro-Wilk test to each of these samples and observed that 90 out of 100 p-values were greater than 0.05, *i.e.*, that the normal approximation seems reasonable even for the smallest n_0 in Table 2. This suggests, for example, that in order to obtain a Bayesian coverage probability of 0.95, we must have $\gamma = 1/1.96^2$. In general, if we want a Bayesian coverage probability approximately equal to $1 - \rho$, we must set $\gamma = 1/[\Phi^{-1}(1 - \rho/2)]^2$, where $\Phi^{-1}(\cdot)$ is the inverse probability function of the standard normal distribution. In other words, larger coverage probabilities requires smaller values for γ , which places more emphasis on the center than on the length of the corresponding credible interval.

We also observe that when the cost of sampling an aliquot c increases, the optimal sample size (and consequently, the average interval length) decreases (increases) under either model, but at the expense of an increase in the total cost (Tables 1-2). For example, if we set the prior variance equal to 4, $\gamma = 1/2$ and $\phi = 1$, from the results in Table 2, it follows that the optimal sample size for $c = 0.001$ is 347, generating a sampling cost of $C(347) = 0.001 \times 347 = 0.347$; the optimal sample size for $c = 0.010$, on the other hand, is 60, generating a sampling cost $C(60) = 0.010 \times 60 = 0.60$, an increase of $\approx 73\%$.

Although the aggregation parameter ϕ represents an important feature related to the heterogeneity of the organism distribution in the ballast water tank, under the total cost minimization approach, the optimal sample size is only slightly affected when ϕ increases (with the other parameters fixed) for $c = 0.01$. As displayed in Table 2, for $\phi \geq 4$ the optimal sample sizes are almost the same for different values of the prior variance. Also, note that for $c = 0.01$, $\phi = 1$ and prior variance equal to 1 we have no entry in Table 2 because there is no associated optimal sample size. This means that the cost of sampling outweighs the cost of decreasing the Bayes risk and it is not worth obtaining aliquots. This was also observed by Etzioni and Kadane (1993) and Islam and Pettit (2014, Table 1).

Such considerations point to a major difficulty of the proposed approach which is the quantification of the “costs” associated to the Bayes risk and to the sampling effort. Although the latter may be objectively calculated in terms of the technical aspects of the actual collection and analysis methods, the former certainly poses a complicated problem since it depends on quantitatively evaluating consequences of declaring a ship compliant or not based on a rule defined in terms of the credible interval. This is certainly a controversial and difficult problem; however, it permeates directly or indirectly, all methods of sample size determination and decision making.

Table 3: Simulated organism counts obtained via the negative binomial model with $\phi = 8$, $w = 1$ and λ fixed as reported.

λ	counts												
7	8	6	5	10	18	2	5	13	3	8	10	4	7
	7	8	11	6	3	3	4	12	3	1	6	2	7
	8	1	11	3	1	5	8	5	2	8	5	5	11
	4	4	3	11	6	2	10	6	6	7	7	8	
10	11	5	22	23	12	4	13	13	3	18	5	10	10
	15	20	27	3	15	4	5	11	11	21	7	3	6
	10	5	9	8	8	5	12	12	2	5	11	9	14
	10	9	10	15	15	10	11	8	7	7	8	6	
13	15	6	21	31	19	17	5	23	7	13	12	25	24
	6	24	12	3	11	7	23	13	5	3	9	16	9
	9	12	11	7	15	16	3	7	15	12	17	13	11
	13	17	20	9	11	8	9	11	8	12	3	13	

For illustrative purposes, we consider a set of hypothetical organism counts to obtain the associated credible interval based on the optimal sample size. Casas-Monroy *et al.* (2020) obtained estimates for ϕ varying from 8 to 22. For $\phi = 8$, the optimal sample size under the negative binomial/gamma model with $\lambda_0 = 10$, prior variance equal to 4 and $c = 0.010$ is $n_o = 51$ (Table 2). We generated 51 observations from a negative binomial model with $\lambda = 7$, $\phi = 8$ and $w = 1$ (see Table 3). Given the generated observations, we drew a sample of size 10,000 from the posterior distribution of λ with a burn-in of 1,000 iterations and a thinning of 10. The corresponding credible interval $[a^*, b^*]$ is $[6.10, 7.05]$. Now, if we generate 51 observations from a negative binomial model with $\lambda = 10$ (Table 3), the corresponding credible interval is $[9.61, 10.89]$. Finally, if we set $\lambda = 13$ to generate the 51 observations, the corresponding interval is $[11.58, 13.04]$. In all cases, the credible interval contains the value of the parameter of interest and lead to correct decisions relatively to compliance v.s. non-compliance with the D-2 standard.

Acknowledgements

The authors appreciate the enlightening comments of two anonymous referees which improved the paper considerably. This research received financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, grants 153526/2014-9 and 304841/2019-6) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant 2013/21728-2), Brazil. This research was also supported by the Fundação para a Ciência e Tecnologia (FCT), Portugal, under Projects UID/MAT/00006/2019 and UID/MAT/00006/2013.

References

- Amaral Turkman, M. A., Paulino, C. D. and Müller, P. (2019). *Computational Bayesian Statistics: An Introduction*. Cambridge: Cambridge University Press.
- Basurko, O. C. and Mesbahi, E. (2011). Statistical representativeness of ballast water sampling. In *Proceedings of the Institution of Mechanical Engineer, Part M*. Journal of Engineering for the Maritime Environment, 183–190.
- Bernardo, J. M. (1997). Statistical inference as a decision problem: the choice of sample size. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 46, 151–153.
- Carney, K. J., Basurko, O. C., Pazouki, K., Marsham, S., Delany, J. E., Desai, D. V., Anil, A. C. and Mesbahi, E. (2013). Difficulties in obtaining representative samples for compliance with the Ballast Water Management Convention. *Marine Pollution Bulletin*, 68, 99–105.
- Casas-Monroy, O., Rajakaruna, H. and Bailey, S. A. (2020). Improving estimation of phytoplankton abundance and distribution in ballast water discharges. *Journal of Applied Phycology*, 32, 1185–1199.
- Costa, E. G., Lopes, R. M. and Singer, J. M. (2015). Implications of heterogeneous distributions of organisms on ballast water sampling. *Marine Pollution Bulletin*, 91, 280–287.
- Costa, E. G., Lopes, R. M. and Singer, J. M. (2016). Sample size for estimating the mean concentration of organisms in ballast water. *Journal of Environmental Management* 180, 433–438.
- Costa, E. G., Paulino, C. D. and Singer, J. M. (2020). ssdet: Sample size determination in frequentist and Bayesian approaches. <http://www.github.com/eliardocosta/ssdet>. R package version 0.0.0.9400.
- Costa, E. G., Paulino, C. D. and Singer, J. M. (2021). Sample size for estimating organism concentration in ballast water: A Bayesian approach. *Brazilian Journal of Probability and Statistics*, 35, 158–171.
- De Santis, F. and Gubbiotti, S. (2017). A decision-theoretic approach to sample size determination under several priors. *Applied Stochastic Models in Business and Industry*, 33, 282–295.
- Etzioni, R. and Kadane, J. B. (1993). Optimal experimental design for another's analysis. *Journal of the American Statistical Association*, 88, 1404–1411.
- Ferguson, T. S. (1996). *A course in large sample theory*. London: Chapman and Hall.
- First, M. R., Robbins-Wamsley, S. H., Riley, S. C., Moser, C. S., Smith, G. E., Tamburri, M. N. and Drake, L. A. (2013). Stratification of living organisms in ballast tanks: how do organism concentrations vary as ballast water is discharged? *Environmental Science and Technology*, 47, 4442–4448.
- Frazier, M., Miller, A. W., LeeII, H. and Reusser, D. A. (2013). Counting at low concentrations: the statistical challenges of verifying ballast water discharge standards. *Ecological Applications*, 23, 339–351.
- Gollasch, S. and David, M. (2017). Recommendations for representative ballast water sampling. *Journal of Sea Research*, 123, 1–15.
- Hastings, W. K. (1970). Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57, 97–109.
- IMO (2004). International convention for the control and management of ship ballast water and sediments. <http://www.imo.org/>.
- Islam, A. F. M. (2011). *Loss functions, utility functions and Bayesian sample size determination*. Ph.D. thesis. Queen Mary, University of London.
- Islam, A. F. M. S. and Pettit, L. I. (2012). Bayesian sample size determination using linex loss and linear cost. *Communications in Statistics-Theory and Methods*, 41, 223–240.
- Islam, A. F. M. S. and Pettit, L. I. (2014). Bayesian sample size determination for the bounded linex loss function. *Journal of Statistical Computation and Simulation*, 84, 1644–1653.
- Lindley, D. V. (1997). The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 129–138.
- Marbuah, G., Gren, I.-M. and McKie, B. (2014). Economics of harmful invasive species: a review. *Diversity*, 6, 500–523.

- McCarthy, S. A., McPhearson, R. M., Guarino, A. and Gaines, J. (1992). Toxigenic *Vibrio cholerae* 01 and cargo ships entering Gulf of Mexico. *Lancet*, 339, 624–625.
- McCullagh, P. and Nelder, J. (1989). *Generalized linear models*, 2nd ed. London: Chapman and Hall.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1092.
- Miller, A. W., Frazier, M., Smith, G. E., Perry, E. S., Ruiz, G. M. and Tamburri, M. N. (2011). Enumerating sparse organisms in ships' ballast water: why counting to 10 is not easy. *Environmental Science and Technology*, 45, 3539–3546.
- Müller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association*, 90, 1322–1330.
- Parmigiani, G. and Inoue, L. Y. T. (2009). *Decision theory: principles and approaches*. New York: John Wiley and Sons.
- Pham-Gia, T. and Turkkan, N. (1992). Sample size determination in Bayesian analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41, 389–392.
- Raiffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. Boston: Wiley Cambridge.
- Rice, K. M., Lumley, T. and Szpiro, A. A. (2008). Trading bias for precision: decision theory for intervals and sets. <http://www.bepress.com/uwbiostat/paper336>. Working Paper 336, UW Biostatistics.
- Sahu, S. K. and Smith, T. M. F. (2006). A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 235–253.
- Schervish, M. (1995). *Theory of Statistics*. New York: Springer-Verlag.
- Strayer, D. L. (2010). Alien species in fresh waters: ecological effects, interactions with other stressors, and prospects for the future. *Freshwater Biology*, 55, 152–174.

Bayesian classification for dating archaeological sites via projectile points

Carmen Armero¹, Gonzalo García-Donato², Joaquín Jiménez-Puerto³,
Salvador Pardo-Gordó³ and Joan Bernabeu³

Abstract

Dating is a key element for archaeologists. We propose a Bayesian approach to provide chronology to sites that have neither radiocarbon dating nor clear stratigraphy and whose only information comes from lithic arrowheads. This classifier is based on the Dirichlet-multinomial inferential process and posterior predictive distributions. The procedure is applied to predict the period of a set of undated sites located in the east of the Iberian Peninsula during the 4th and 3rd millennium cal BC.

MSC: 62F15, 62H30, 01A10.

Keywords: Bifacial flint arrowheads, chronological model, Dirichlet-multinomial process, posterior predictive distribution, radiocarbon dating.

1 Introduction

Dating is a key element for archaeologists. A time scale to locate the information collected from excavations and field work is always necessary in order to build, albeit with uncertainty, our most remote past. Archaeological scientists generally use stratigraphic expert information and dating techniques for examining the age of the relevant artifacts. Bayesian inference is commonly used in archaeology as a tool to construct robust chronological models based on information from scientific data as well as expert knowledge (e.g. stratigraphy) (Buck, Cavanagh and Litton, 1996).

Radiocarbon dating is one of the most popular techniques for obtaining data due to carbon's presence in any being that has lived on Earth. However, it is not always possible in all studies to collect organic material and obtain that type of information or to have good stratigraphic references. In these cases, the challenge is to be able to assign non

¹ Departament d'Estadística i IO. Universitat de València, Carrer Doctor Moliner 50, 46100, Burjassot, Spain.

² Department of Economics and Finance. Universidad Castilla-La Mancha. Edificio Jurídico-Empresarial "Melchor de Macanaz", Plaza de la Universidad, 1, 02071 Albacete, Spain.

³ Department of Prehistory, Archaeology and Ancient History, Universitat de València, Avda. Blasco Ibáñez 28, 46010 València, Spain.

Received: November 2020

Accepted: March 2021

radiocarbon dated collections to specific chronological times. The relevant information is based on cultural material that includes elements with markers that point out the different cultural traits of the social groups involved as well as the social relationships between them. One of these useful items is the lithic productions, and more specifically the arrowheads.

During the 4th and 3rd millennium cal BC bifacial flint arrowheads appear and spread in the east of the Iberian Peninsula. Archaeological research suggests that the shape of these arrowheads could be related with specific period and/or geographical social units spatially defined.

In this context, we propose an automatic Bayesian procedure, very popular in text classification (Wang, Hodges and Tang, 2003), based on predictive probability distributions for classifying the period to which an undated site belongs according to the type and number of arrows found in it. This proposal takes into account the Dirichlet-multinomial inferential process for learning about the proportion of different types of arrowheads in each chronological period, and the concept of posterior predictive distribution for a new undated site.

This paper is organized in five sections. Following this introduction, Section 2 briefly introduces the archaeological framework and the lithic material that will be the basis for the classification process. Section 3 describes the two stages of the Bayesian statistical analysis. The first is of an inferential type and focuses on the study of the abundance of different types of arrows in the different periods considered. The second uses the information from the first stage to predict the period of an undated site from the number and type of arrowheads encountered. Section 4 applies the methodological procedure from the previous section to a set of sites in the east of the Iberian Peninsula during Late Neolithic and Chalcolithic (4th-3rd millennium BC). Finally, Section 5 concludes.

2 Chronological periods and lithic information

One of the main goals in archaeological research is focused on the way the members of the prehistoric cultures interact with the landscape and the objects. From an evolutive perspective, the way human cultures change through space-time is determined by inheritance patterns, adaptation and interaction (Shennan, Crema, and Kerig, 2015). Therefore, the analysis of items from the archaeological records, able to capture the cultural evolution of the human groups, would be a main goal for the researcher.

The concept of “culture” covers many factors. Hence, we will use the material culture as an archaeologic proxy in order to analyse the evolution and dispersion of the cultural traits in the study area. Not all the items included in material culture are useful for that. Those which show a wide geographic and cultural dispersion or whose variability is low are not convenient to detect changes. This is not the case with lithic productions, and more specifically arrowheads, which provide information not only for understanding

the socio-economic and cultural structures of human groups, but they can be used as a valuable tool for chronological dating.

The arrival of the neolithic economy, based on domestic resources, in the Iberian Peninsula is dated on the first half of the 6th millennium cal BC. We will have to wait until the 4th-3rd millennium to be able to witness clear winds of change. This is the moment of the appearance of a higher level of hierarchy in some societies. The Late Neolithic (4th-3rd millennium cal BC) in the oriental Iberian façade is the time of the transit to a higher complexity in social and economic terms. This process will last long and it will crystallize by the end of the 3rd millenium cal BC (Bernabeu and Orozco, 2014). The evaluation of this process in such a huge frame faces some problems which need to be addressed. One of these difficulties is closely associated with the chronological attribution of a big part of the period's archaeological record due to scarce radiocarbon data.








Type 1	with rhomboid or rhombus-eye shape	
Type 2	with side appendages or cruciform	
Type 3	leaf-like	
Type 4	with peduncle but without flints	
Type 5	with a concave base	
Type 6	asymmetric	
Type 7	with peduncle and flints	

Figure 1: Arrowhead types used for the study.

The classification of the arrowheads in this period is based on the previous works performed around the typological formalization for the study area. They are mainly inspired

by morpho-descriptive typologies. Therefore, the classification contains a functional and morphological meaning. Arrowheads constitute a very representative tool group of the Late Neolithic and Chalcolithic. Their function is quite proved thanks to the studies in traceology, experimental archaeology and etnoarchaeology. Some well known examples are the spectacular findings of arrowheads still nailed into the victim bones, present in many burials from the 4th and 3rd millennium BC (i.e. San Juan ante Portam Latinam: Vegas 2007). We cannot forget the awesome finding of a full equipment Ötzi, the “Iceman”, discovered in the Alps (Cave-Browne, 2016), and exceptionally conserved. Moreover, the existence of excavated sites (Ereta del Pedregal) in which the whole arrowhead operative chain process can be observed, has provided additional information (Juan-Cabanilles, 1994).

The arrowhead types present in the archaeological records have been classified in seven types following a morphological criterion, based on previous typologies for the study area (Juan-Cabanilles, 2008) (See Figure 1).

3 Bayesian classification process

Bayesian classification within the framework of archaeological datation with lithic information will provide a probability distribution for the period to which an undated site belongs in which a given set of different types of arrowheads has been found. This probability distribution depends on the knowledge of the abundance of each type of arrowheads in each period, expressed via the posterior distribution for the probability associated with each type of arrowhead, and the posterior predictive distribution for the period of that particular updated site.

3.1 Dirichlet-multinomial inferential process

Let Y_{ij} be the random variable that describes the number of type j , $j = 1, \dots, J$ arrowheads, of the total n_i collected in the sites belonging to period i , $i = 1, \dots, I$. We define the random vector $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{i,J-1})^\top$ and the probability vector $\theta_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{i,J-1})^\top$, where θ_{ij} is the probability that an arrowhead of period i is of type j . A probabilistic model for $Y_i | \theta_i$ is the multinomial distribution, $\text{Mn}(\theta_i, n_i)$, with probability distribution

$$f(y_i | \theta_i) = \frac{n_i!}{\left(\prod_{j=1}^{J-1} y_{ij}!\right)} \left(\prod_{j=1}^{J-1} \theta_{ij}^{y_{ij}}\right) \theta_{iJ}^{y_{iJ}}, \quad (1)$$

where y_i is an observation of Y_i , $y_{iJ} = n_i - \sum_{j=1}^{J-1} y_{ij}$ is the total number of arrowheads of type J in the sites of period i , and $\theta_{iJ} = 1 - \sum_{j=1}^{J-1} \theta_{ij}$ is the probability that an arrowhead of period i is of type J .

The combination of a multinomial sampling model with a conjugate Dirichlet prior distribution was proposed by Lindley (1964) and Good (1967) as the generalisation of the beta-binomial model. The Dirichlet distribution for θ_i with parameters $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iJ})^\top$, $\alpha_{ij} > 0, j = 1, \dots, J$, $\text{Dir}(\alpha_i)$, is a multivariate continuous distribution with joint density function

$$\pi(\theta_i) = \frac{\Gamma(\alpha_{i+})}{\prod_{j=1}^J \Gamma(\alpha_{ij})} \left(\prod_{j=1}^{J-1} \theta_{ij}^{\alpha_{ij}-1} \right) \theta_{iJ}^{\alpha_{iJ}-1}, \quad (2)$$

where $\Gamma(\cdot)$ represents the gamma function and $\alpha_{i+} = \sum_{j=1}^J \alpha_{ij}$.

We assume an inferential process for each $\theta_i, i = 1, \dots, I$ in the framework of the Dirichlet-multinomial process with a non-informative prior distribution for θ_i that gives all the protagonism of the process to the data. There are many proposals for elicit the parameters α_i in a non-informative way: Haldane's prior, Perks' prior or reference distance prior, hierarchical approach prior and Jeffreys' prior or common reference prior, and Bayes-Laplace prior. All them have good theoretical properties but they also have some small shortcomings. We choose the Perks' prior as a result of Alvares, Armero and Forte (2018). This prior was firstly proposed by Perks (1947), but recently it has been also obtained as the reference distance prior by Berger et al. (2015). This is a Dirichlet distribution with all parameters equal to $1/J$, where J is the number of arrow types. Figure 2 shows the density and other characteristics of a Perk's distribution with three categories.

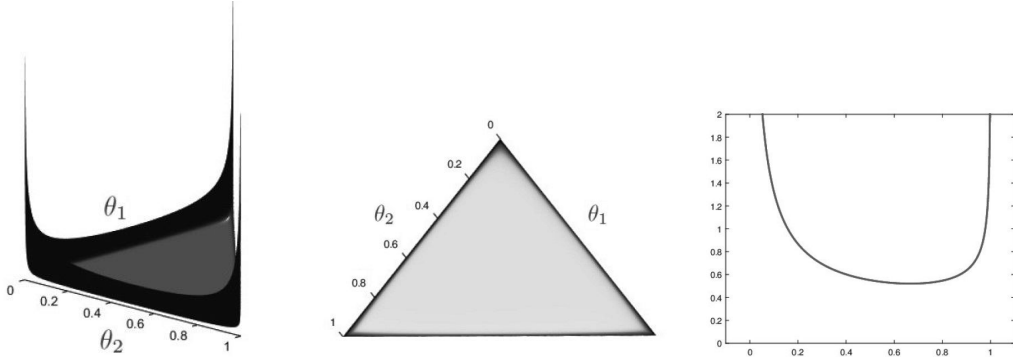


Figure 2: Perks' distribution when the number of types of arrowheads is $J = 3$ (a), its projection onto the simplex triangle (b), and the marginal prior distribution for each individual component, a beta distribution with parameters $1/3$ and $2/3$, $\text{Be}(1/3, 2/3)$, which maintains high density values close to 0 and 1 (c).

The posterior distribution for θ_i when data y_i are observed is also a Dirichlet distribution (Lindley, 1964; Good, 1967),

$$\pi(\theta_i | y_i) = \text{Dir}(\alpha_{i1} = y_{i1} + (1/J), \dots, \alpha_{iJ} = y_{iJ} + (1/J)). \quad (3)$$

This posterior distribution has an important and positive feature: never assigns absolute probabilities 1 or 0 to the presence of any type of headarrows. This fact avoids working with absolute values of the probabilities, 0 and 1, which would prevent future updates of their values generated by new data.

The marginal posterior distribution for each probability θ_{ij} is the beta distribution (Gelman *et al.*, 2014)

$$\pi(\theta_{ij} | y_i) = \text{Be}(\alpha_{ij}, \alpha_{i+} - \alpha_{ij}), \quad (4)$$

with posterior mean and variance α_{ij}/α_{i+} and $\alpha_{ij}(\alpha_{i+} - \alpha_{ij})/(\alpha_{i+}^2(\alpha_{i+} + 1))$, respectively.

3.2 Predictive process

After learning about the distribution of the proportion of arrowheads types in each site, we have to assign a probability distribution to the random variable that describes the period m^* to which a new undated site s^* belongs given that a total of n^* arrowheads $y^* = (y_1^*, \dots, y_J^*)^T$ have been observed in it. Following Bayes' theorem:

$$P(m^* = m_i | y^*, y) \propto P(Y^* = y^* | m^* = m_i, y) P(m^* = m_i | y), \quad i = 1, \dots, I, \quad (5)$$

where $y = (y_1, \dots, y_I)^T$ are the observed data in the previous estimation process and $Y^* = (Y_1^*, \dots, Y_J^*)^T$ is the random vector that describes the number of arrowheads of the different types that will be recorded in that new site. It is important to note that Y and Y^* in capital letters refer to the random vector that generate or will generate the data y and y^* , respectively, which we always represent by lower case letters. The asterisk is used to represent the subsequent random variables and observations of the prediction process.

The posterior predictive distribution in (5) is proportional to the product of two terms. The first one is:

$$\begin{aligned} P(Y^* = y^* | m^* = m_i, y) &= \int P(Y^* = y^* | \theta_i, m^* = m_i, y) \pi(\theta_i | m^* = m_i, y) d\theta_i \\ &= \int \frac{n^*!}{y_1^*! y_2^*! \dots y_J^*!} \theta_{i1}^{y_1^*} \theta_{i2}^{y_2^*} \dots \theta_{iJ}^{y_J^*} \frac{\Gamma(\alpha_{i+})}{\prod_{j=1}^J \Gamma(\alpha_{ij})} \theta_{i1}^{\alpha_{i1}-1} \theta_{i2}^{\alpha_{i2}-1} \dots \theta_{iJ}^{\alpha_{iJ}-1} d\theta_i \\ &= \frac{n^*!}{y_1^*! y_2^*! \dots y_J^*!} \frac{\Gamma(\alpha_{i+})}{\prod_{j=1}^J \Gamma(\alpha_{ij})} \int \theta_{i1}^{\alpha_{i1}+y_1^*-1} \theta_{i2}^{\alpha_{i2}+y_2^*-1} \dots \theta_{iJ}^{\alpha_{iJ}+y_J^*-1} d\theta_i \\ &= \frac{n^*!}{y_1^*! y_2^*! \dots y_J^*!} \frac{\Gamma(\alpha_{i+})}{\Gamma(\alpha_{i+} + n^*)} \prod_{j=1}^J \frac{\Gamma(\alpha_{ij} + y_j^*)}{\Gamma(\alpha_{ij})}. \end{aligned}$$

The first probability in the integrand, $P(Y^* = y^* | \theta_i, m^* = m_i, y)$, is associated with new experimental results in the presence of θ_i and the data y from the estimation process

which are irrelevant due to the presence of θ_i . It is a multinomial probability computed from (1). The second term, $\pi(\theta_i | m^* = m_i, y)$, is the Dirichlet posterior distribution for θ_i given in (3).

The second element in the product in (5), $P(m^* = m_i | y)$, can be estimated as the proportion of sites in the sample for each of the periods under consideration (Barber, 2012).

4 East of the Iberian Peninsula sites during the 4th and 3rd millennium cal BC

We apply the classification procedure above to a set of undated sites in the East of the Iberian Peninsula during the 4th and 3rd millennium cal BC. Data for the inferential process of the study come from 31 archaeological sites radiocarbon dated with arrowheads, clear contexts and stratigraphy.

4.1 Inferential process

All 14C dated sites have been filtered using only those whose radiocarbon dates come from short-lived singular samples. The final levels used for the periodization are: Arenal de la Costa (Bernabeu, 1993), Barranc del Migdia (Soler Díaz et al., 2016), Beniteixir (Pascual Beneyto, 2010), Camí de Missena (Pascual Beneyto, Barberà and Ribera, 2005), Colata (Gómez Puche et al., 2004), Cova del Randero (Soler Díaz et al., 2016), Cova dels Diablets (Aguilella, Olaria Puyoles and Gusi Jener, 1999), Jovades (Bernabeu, 1993), La Vital (Pérez-Jordà et al., 2011), Niuat (Bernabeu, Pascual Benito, Orozco Köhler, Badal García, Fumanal García and García Puchol, 1994), and Quintaret (García Puchol et al., 2014). These sites are located in the eastern Mediterranean area. Figure 3

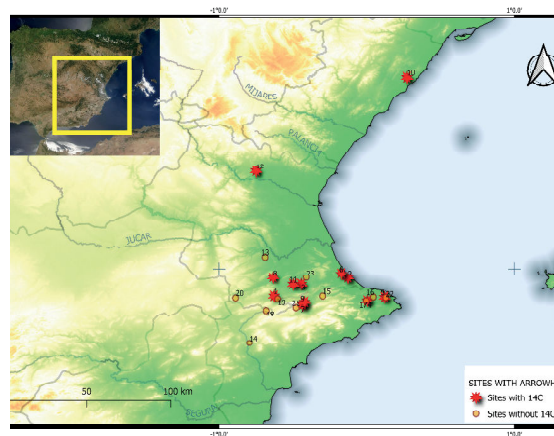


Figure 3: Situation map of the sites with arrowheads present in the study area.

shows a map with the dated sites as well as the sites without ^{14}C datation whose chronological classification is the final object of this study.

Based on the chrono-stratigraphic and available expert information, we have proposed five intervals or chronological periods organization comprised between ca. 4600-3200 cal BC. Table 1 includes the period of each of the periods considered as well as the sites included in each of them.

Each site usually contains many different archaeological levels attached to different moments of occupation. In this specific case, archaeological contexts containing arrowheads have been dated through radiocarbon determinations. Some of these sites contain different dated levels in which arrowheads were present. Hence we have described them with the name of the site and a number to differentiate them. Based on the chrono-stratigraphic and available expert information, we have proposed five successive intervals or chronological periods comprised between ca. 4600-2150 cal BC. These periods have resulted from the application of Bayesian radiocarbon modeling methods to the archaeological information available for each period.

Table 1: Periods and sites extracted from clear archaeological contexts with radiocarbon determinations.

Sites ^{14}C dated	Period
Jovades 1, Jovades 2, and Niuet 1	1
Colata 1, Colata 2, Jovades 3, Jovades 4, Niuet 2, and Quintaret	2
Beniteixir, Diablets 1, Diablets 2, Diablets 3, Jovades 5, La Vital 1, La Vital 2, Migdia 1, Missena 1, Niuet 3, Niuet 4, Randero 1, and Randero 2	3
La Vital 3, Migdia 2, Missena 2, and Missena 3	4
Arenal Costa, La Vital 3, Missena 4, Missena 5, and Missena 6	5

Table 2: Posterior Dirichlet distribution for the proportion of arrowheads from type 1 to type 7 in each of the periods considered.

Period	Posterior distribution
1	Dir(15/7, 22/7, 8/7, 1/7, 1/7, 1/7, 1/7)
2	Dir(29/7, 36/7, 15/7, 8/7, 1/7, 1/7, 1/7)
3	Dir(43/7, 1/7, 43/7, 64/7, 29/7, 1/7, 71/7)
4	Dir(15/7, 1/7, 15/7, 8/7, 15/7, 1/7, 43/7)
5	Dir(1/7, 1/7, 1/7, 15/7, 1/7, 8/7, 36/7)

Table 2 includes the posterior distribution of the different types of arrowheads in each of the periods considered. In all of them the selected prior distribution is the Perk

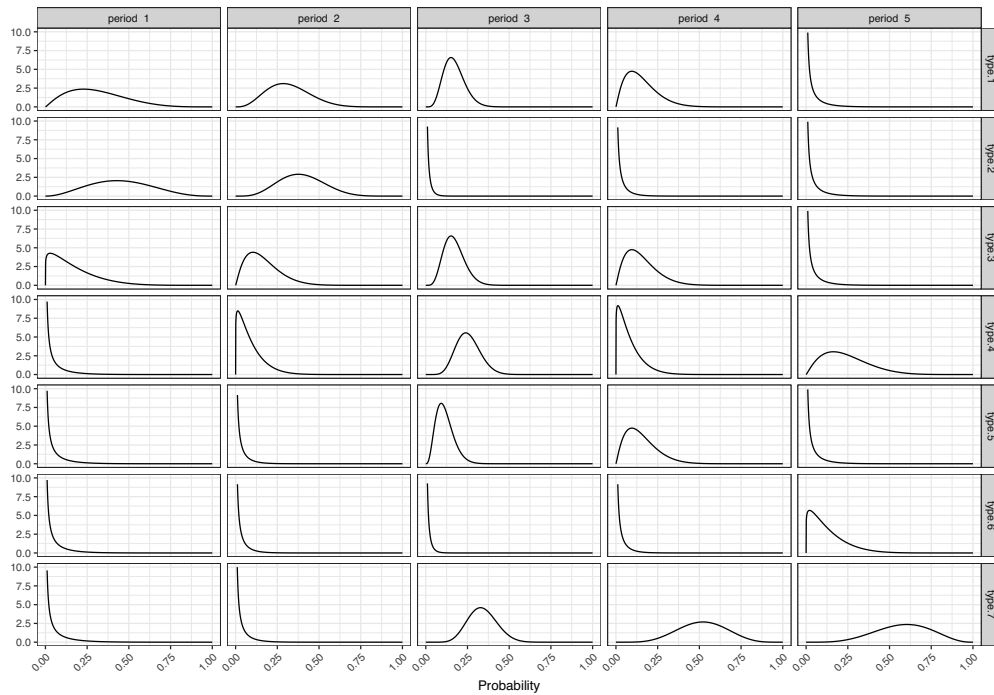


Figure 4: Posterior marginal distribution for the probability associated with each type of arrowhead in each of the periods in the study.

distribution $\text{Dir}(1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7)$. Therefore, those parameters of the corresponding posterior distribution that continue to be worth $1/7$ correspond to those types of arrows that have not been observed in the sample.

Table 3 shows the posterior mean for the probability associated with each type of arrowhead in each of the periods in the study. Figure 4 shows the posterior marginal distribution of the probability of the different types of arrowheads in each of the five chronological periods considered. Results in Table 3 and Figure 4 indicate that the distribution of the different types of arrowheads is very similar in Periods 1 and 2: Type 1 and 2 arrowheads are the most abundant and about the 75% and 70% of the total of arrowheads in both periods are type 1 or 2. Type 3 arrowheads have poor relevance in both Periods and types 4, 5, 6, and 7 are virtually nonexistent. In Period 3, we find practically no type 2 and 6 arrowheads. The remaining arrowheads in this period have a presence quite similar but type 4 and 7 have a slightly higher presence. Period 4 shows a large presence of type 7 arrows and, to a lesser extent, of type 1, 3 and 5 arrows (probabilities of about 0.15). Arrowheads of type 2 and 6 have no relevance. Approximately 57% and 24% of the arrows of Period 5 are of type 7 and 4, respectively. The remaining arrowhead types, except possibly those of type 6, are essentially irrelevant.

Table 3: Posterior mean of the probability associated to each type of arrowhead in each of the periods of the study.

Type	Period 1	Period 2	Period 3	Period 4	Period 5
1	0.3061	0.3187	0.1706	0.1531	0.0159
2	0.4490	0.3956	0.0040	0.0102	0.0159
3	0.1633	0.1648	0.1706	0.1531	0.0159
4	0.0204	0.0879	0.2540	0.0816	0.2380
5	0.0204	0.0110	0.1151	0.1531	0.0159
6	0.0204	0.0110	0.0040	0.0102	0.1270
7	0.0204	0.0110	0.2817	0.4387	0.5714

4.2 Predictive process

Undated sites between the 4th and 3rd millennium cal BC. used to explore the predictive approach include burial sites, villages, and caves: Barranc Cafer 2, Barranc Parra 3, Casa Colorà, Cova Ampla del Montgó, Cova Santa Vallada B, Cova de les Aranyes, Cova dels Anells, Cova del Negre, Cova del Petrolí, Cova Pardo, Cova Santa Vallada A, Ereta I, Ereta II, Ereta III, Ereta IV, Escurrupeña, Font de Mahiques, Garrofer 3, Garrofer K, Garrofer I-J, Rambla Castellarda, Sima de la Pedrera, Niuets, Torreta UE1, and Torreta UE2 (See Figure 3).

The posterior probability that a new site belongs to each of the periods considered was estimated as 0.15 for Periods 1, 4 and 5, 0.20 for Period 2, and 0.35 for Period 3.

Figure 5 presents the posterior predictive distribution of the period to which the above undated sites belong, whose only available information is based on the number and type of arrows found collected.

The results obtained show a high concordance with the expert information provided by archaeologists. Thus, for example, in those sites that present stratigraphic correlations (Ereta del Pedregal and La Torreta) the chronological evaluation obtained from the predictive approach is consistent with the chrono-statigraphical information. The case of Cova Santa de Vallada B is interesting, which from the archaeological information is situated in phase 3-4. However, based on Bayesian modeling, this indicates that it should be located in Period 3. This aspect is totally coherent not only because of the typology of the arrowheads themselves but also because of the presence of other diagnostic elements such as the presence of metal and the absence of bell-beaker ceramics. The result is totally consistent with the cases of Casa Colorà and Cova del Garrofer I-J, which both the previous experience and the Bayesian application place in Period 3. Finally, there are some cases in which the results qualify the chronological proposal established by expert knowledge, such as the case of Barranc de Parra 3, where previous knowledge places it in Period 2-3 but predictive analysis places it either in Period 1 or in Period 4. In this sense, we must bear in mind both that there may be a persistence of certain types

of arrowheads throughout the entire sequence analyzed, as is the case of the arrowheads of the peduncle, as well as the possible reuse of projectiles located in places of habitat as has been documented in the Clovis culture, North America. In this sense both the incorporation of other complementary diagnostic archaeological information (presence of metal and bell-shaped ceramics) may help to establish a more precise chronology.

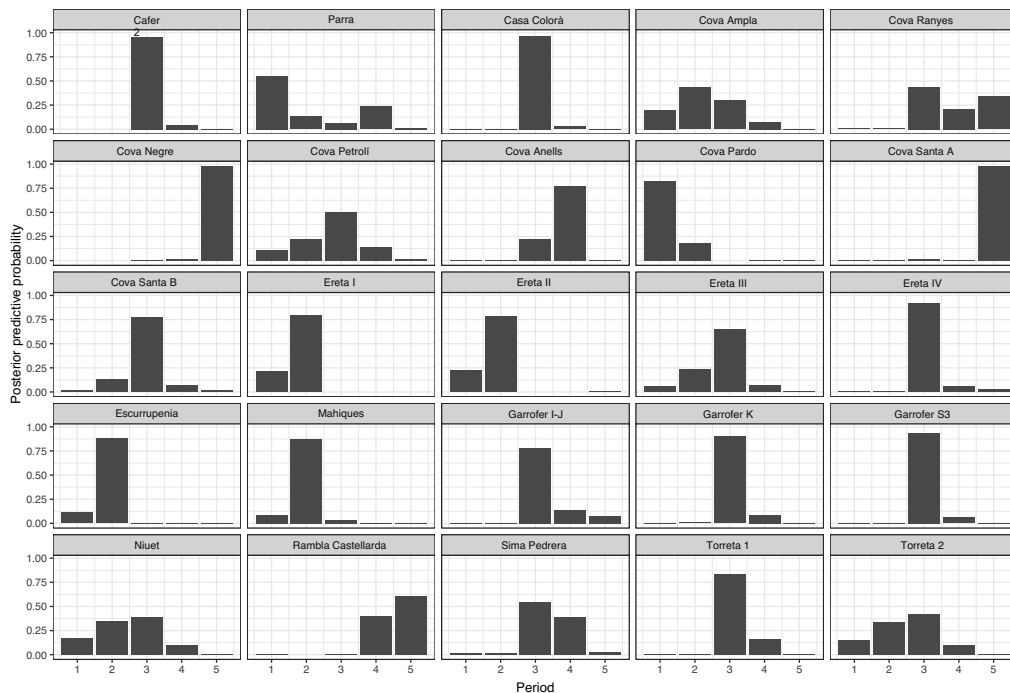


Figure 5: Posterior predictive distribution associated with each chronological period for each non-dated site in the study.

Conclusions

In short, results obtained present a good agreement with the expert information of the archaeologists, so it is a proposal that can be very useful in archaeological research. However, there is no doubt that both the application of stratigraphic contexts of higher resolution and the use of associated radiometric dates related to the most diagnostic archaeological items will allow to improve this approach.

Acknowledgements

This paper has been partially supported by grants PID2019-106341GB-I00 and FPU16/00781 from the Ministerio de Ciencia e Innovación (MCI, Spain), and grant AICO/2018/005 from Generalitat Valenciana. JJP is supported by grant FPU16/00781 from the Mi-

nisterio de Ciencia e Innovación and SPG by Generalitat Valenciana postdoctoral grant APOST-2019/179. The authors wish to thank the referees for helpful suggestions in improving the paper and are very grateful to Danilo Alvares for his help with some of the graphics in the paper.

References

- Aguilella, G.C., Olaria Puyoles, R. and Gusi Jener, F. (1999). El jaciment prehistòric de la Cova dels Diablets (Alcalà de Xivert, Castelló). *Quaderns de Prehistòria i Arqueologia de Castelló*, 20, 7-36.
- Alvares, D., Armero, C. and Forte, A. (2018). What does objective mean in a Dirichlet-multinomial process? *International Statistical Review*, 86, 106-118.
- Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press
- Bernabeu J. (1993). El III milenio aC en el País Valenciano: los poblados de Jovades (Cocentaina, Alacant) y Arenal de la Costa (Ontinyent, València). *SAGVNTVM. Papeles del Laboratorio de Arqueología de València*, 26, 9-179.
- Bernabeu, J., Pascual Benito, J.L., Orozco Köhler, T., Badal García, E., Fumanal García, M.P. and García Puchol, O. (1994). Niuet (L'Alqueria d' Asnar). Poblado del III Milenio aC. *Recerques del Museu d'Alcoi*, 3, 9-74.
- Buck, I.C.E., Cavanagh, W.G. and Litton, C.D. (1996). *Bayesian Approach to Interpreting Archaeological Data*. Chichester: Wiley.
- García Puchol, O., Molina Balaguer, L., Cotino Villa, F., Pascual Benito, J.L., Orozco Köhler, T., Pardo Gordó, S., Carrión Marco, Y., Pérez-Jordà, G., Clausí Sifre, M. and Gimeno Martínez, L. (2014). Hábitat, marco radiométrico y producción artesanal durante el final del Neolítico y el Horizonte Campaniforme en el corredor de Montesa (València). Los yacimientos de Quintaret y Corcot. *Archivo de Prehistoria Levantina*, 30, 159-211.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2014). *Bayesian Data Analysis*. Third Edition. Boca Raton: Chapman and Hall.
- Gómez Puche, M., Díez Castillo, A., Verdasco, C., García Borja, P., Mclure, S., López Gila, M.D., García Puchol, O., Orozco, Köhler, T., Pascual Benito, J. and Carrión Marco, Y. (2004). El yacimiento de Colata (Montaverner, València) y los poblados de silos del IV milenio en las comarcas centro-meridionales del País Valenciano. *Recerques del Museu d'Alcoi*, 13, 53-127.
- Good, I.J. (1967). A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society, Series B*, 29, 399-431.
- Lindley, D.V. (1964). The Bayesian analysis of contingency tables. *Annals of Mathematical Statistics*, 35, 1622-1643.
- Pascual Beneyto, J., Barberà, M. and Ribera, A. (2005). El camí de Missena (La Pobla del Duc). Un interesante yacimiento del III milenio en el País Valenciano. *Actas del III Congreso de Neolítico en la Península Ibérica*, 803-814.
- Pascual Beneyto, J. (2010). El Barranc de Beniteixir. En *Restos de vida, restos de muerte: la muerte en la Prehistoria*. Exposición celebrada en el Museu de Prehistòria de València del 4 de febrero al 30 de mayo de 2010, 191-194. Museu de Prehistòria de València.
- Pérez-Jordà, G., Bernabeu Aubán, J., Carrión, Y., García Puchol, O., Molina Balaguer, L. and Gómez Puche, M. (2011). La Vital (Gandía, València). Vida y muerte en la desembocadura del Serpis durante el III y el I milenio AC. *Serie de Trabajos Varios del Servicio de Investigación Prehistórica*.
- Soler Díaz, J.A., Roca de Togores Muñoz, C., Esquembre-Bebíá, M.A., Gómez Pérez, O., Boronat Soler, J.D., Benito Iborra, M., Ferrer García, C. and Bolufer Marqués, J. (2016). Progresos en la investigación

del fenómeno de inhumación múltiple en La Marina Alta (Alicante): A propósito de los trabajos desarrollados en la Cova del Randero de Pedreguer y en la Cova del Barranc del Migdia de Xàbia. En *Del Neolític a l'edat de bronze en el Mediterrani occidental*, 1.a ed., 323-348. Museu de Prehistòria de València.

Wang, Y., Hodges, J. and Tang, B. (2003). Classification of web documents using a naive Bayes method. *15th IEEE International Conference on Tools with Artificial Intelligence*, 124, 560-564.

Joint outlier detection and variable selection using discrete optimization

Mahdi Jammal¹, Stephane Canu² and Maher Abdallah³

Abstract

In regression, the quality of estimators is known to be very sensitive to the presence of spurious variables and outliers. Unfortunately, this is a frequent situation when dealing with real data. To handle outlier proneness and achieve variable selection, we propose a robust method performing the outright rejection of discordant observations together with the selection of relevant variables. A natural way to define the corresponding optimization problem is to use the ℓ_0 norm and recast it as a mixed integer optimization problem. To retrieve this global solution more efficiently, we suggest the use of additional constraints as well as a clever initialization. To this end, an efficient and scalable non-convex proximal alternate algorithm is introduced. An empirical comparison between the ℓ_0 norm approach and its ℓ_1 relaxation is presented as well. Results on both synthetic and real data sets provided that the mixed integer programming approach and its discrete first order warm start provide high quality solutions.

MSC: 62J05, 62J20, 62J07, 62G35, 90C11, 68T05.

Keywords: Robust optimization, statistical learning, linear regression, variable selection, outlier detection, mixed integer programming.

1 Introduction

We consider the linear regression model:

$$y = X\beta + \epsilon.$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the model matrix, $\beta \in \mathbb{R}^p$ is the vector of regression coefficients and $\epsilon \in \mathbb{R}^n$ is the error vector. It is convenient to estimate β with a sparse vector, especially for high values of p .

¹ Institut National des Sciences Appliquees (INSA) de Rouen 685 Avenue de l'Universite 76800 Saint-Etienne du Rouvray; Lebanese Univerity, Beirut, Lebanon.

² Institut National des Sciences Appliquees (INSA) de Rouen 685 Avenue de l'Universite 76800 Saint-Etienne du Rouvray.

³ Lebanese University, Faculty of public health, Hadath, Beirut, Lebanon.

Received: October 2020

Accepted: April 2021

It is well known that dimension reduction or feature selection is an effective strategy to handle contaminated data and to deal with high dimensionality while providing better prediction (Bertsimas, King and Mazumder, 2015). Outliers, i.e. atypical or corrupted observations, can also have a considerable bad influence on estimators (Yang et al., 2010; Rousseeuw and Hubert, 2018). Usually, outliers are eliminated in a time consuming data cleaning pretreatment (Hodge and Austin, 2004; Campos et al., 2016) while variable selection is performed together with parameter estimation using the Lasso (Tibshirani, 1996), its variants (Tibshirani, Wainwright and Hastie, 2015) or the best subset (Bertsimas et al., 2015) algorithms just to name a few. For a recent comparison of these algorithms, see for instance Hastie, Tibshirani and Tibshirani (2017). However, it is well known that, due to the ordinary least square (OLS) criterion used in the lasso, it is not robust to outliers. For instance, Alfons et al. (2013) show that the breakdown point of the lasso is $1/n$, that is, only one single outlier can make the lasso estimate completely unreliable.

Different attempts have been made to solve this problem by mixing variable selection and outlier detection. A popular idea is to replace the OLS criterion of the lasso by a loss robust to outliers such as the absolute deviation (Wang, Li and Jiang, 2007), the least trimmed squares estimator (Alfons et al., 2013) introduced by Rousseeuw and Leroy (1987) or the Huber's loss (Dalalyan and Thompson, 2019). Also, to deal with the specific case of cellwise contamination, that is the presence of outliers in the design matrix, Öllerer, Alfons and Croux (2016) introduced the shooting S-estimator.

However, none of these approaches considered the use of the pseudo ℓ_0 norm as recently introduced by Bertsimas et al. (2015). In this paper we propose to get robust estimates by solving these two problems of variable selection and outliers detection together using pseudo ℓ_0 norms for both. Such an approach leads to reformulating the double robust regression problem as a mixed integer program providing a global solution with convergence guarantee in case of early stopping as well as flexibility and adaptability. It also allows the use of efficient solvers such as Gurobi, the one used in our experiments to obtain good results on both synthetic and real data.

Brief Context and Background

Let $X = (x_1, \dots, x_n)^\top$ be a $n \times p$ design matrix and $y \in \mathbb{R}^n$ a response vector. We consider the following linear model to accommodate outliers:

$$\forall i \in \{1, \dots, n\}, \quad y_i = \begin{cases} x_i^\top \beta + \epsilon_i & \text{if observation } i \text{ is regular} \\ \gamma_i & \text{if observation } i \text{ is an outlier to be trimmed,} \end{cases} \quad (1)$$

where $\beta \in \mathbb{R}^p$ is the unknown parameter vector to be estimated, $\epsilon \in \mathbb{R}^n$ is the noise vector and $\gamma \in \mathbb{R}^n$ an intervention vector. A way to model doubtful observations to be trimmed is to introduce a vector $\tau \in \mathbb{R}^n$ modeling outliers:

$$\forall i \in \{1, \dots, n\}, \quad \tau_i = \begin{cases} 0 & \text{if observation } i \text{ has to be taken into account} \\ y_i - x_i^\top \beta - \epsilon_i & \text{if observation } i \text{ is an outlier to be trimmed,} \end{cases}$$

The model (1) can be rewritten as the following linear model She and Owen (2011):

$$y = X\beta + \epsilon + \tau. \quad (2)$$

We are interested in minimizing the norm of the noise vector while selecting k_v variables and removing k_o outliers, that is, solving the following optimization problem Chen, Caramanis and Mannor (2013), for some $q \in \{1, 2\}$,

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p, \tau \in \mathbb{R}^n} \quad & \frac{1}{q} \|X\beta + \tau - y\|_q^q \\ \text{s.t.} \quad & \|\beta\|_0 \leq k_v \\ & \|\tau\|_0 \leq k_o, \end{aligned} \quad (3)$$

This formulation allows the selection of relevant variables and the avoidance of outliers. When $k_o = 0$, no outlier detection is performed and this problem boils down to the best subset selection problem Miller (2002); Bertsimas et al. (2015); Miyashiro and Takano (2015). When $k_v = p$, no variable selection is performed, the resulting problem is known as the least trimmed squares regression problem Rousseeuw and Leroy (1987); Giloni and Padberg (2002). Due to the nature of the cardinality constraints, Problem (3) is a non-convex optimization problem and has been shown to be NP-hard and considered as an intractable problem. Mainstream research focused on solving a relaxed version of Problem (3), by using the ℓ_1 norm instead of the ℓ_0 norm:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p, \tau \in \mathbb{R}^n} \quad & \frac{1}{2} \|X\beta + \tau - y\|_2^2 \\ \text{s.t.} \quad & \|\beta\|_1 \leq \lambda \\ & \|\tau\|_1 \leq \gamma \end{aligned} \quad (4)$$

where λ and γ are two nonnegative regularization parameters. Problem (4) will be denoted by ℓ_1 -RR. However, this approach is not globally optimal in the sense of (3) since it will not necessarily provide the same solution provided by (3). We recall that the lagrangian relaxation of Problem (4) is given by:

$$\min_{\beta \in \mathbb{R}^p, \tau \in \mathbb{R}^n} \quad \frac{1}{2} \|X\beta + \tau - y\|_2^2 + \lambda \|\beta\|_1 + \gamma \|\tau\|_1 \quad (5)$$

Statistical properties of Problem (5) have been explored in Dalalyan and Thompson (2019); Nguyen and Tran (2013). To retrieve the global minimum of Problem (3), we propose to recast Problem (3) as a mixed integer optimization problem (MIO), which allows the use of efficient solvers to solve it, “Gurobi” for example. The MIO approach has a computational cost, but two decades of progress enabled its effective practical use for moderately sized problems. We also present a discrete first order algorithm that pro-

vides a high quality solution that could be used as a warm start for the MIO algorithm. In addition, it is useful for high-dimensional data sets since it provides solutions in a short time.

The remainder of the paper is organized as follows. In Section 2, we present our approach for variable selection and outliers detection using the ℓ_0 together with its formulation as a mixed integer optimization allowing to obtain the global solution. Section 3 introduces a relaxation that provides efficiently a local solution to this problem. This is followed by Sections 4 and 5 reporting empirical evidence on both synthetic and real data sets respectively. Finally, the paper is concluded in Section 6.

2 Variable Selection and Outlier Detection as a MIO

We propose to reformulate Problem (3) as a mixed integer (binary) optimization (MIO) problem by introducing binary variables representing whether or not variables and observations are useful.

2.1 Introducing Binary Variables

Variable selection involves the ℓ_0 norm function to count the number of useful variables. This counting function can be represented by introducing p binary variables $z_j \in \{0, 1\}$ such that

$$\|\beta\|_0 = \sum_{j=1}^p z_j \quad \text{and} \quad z_j = 0 \Rightarrow \beta_j = 0.$$

Different approaches can be used to force $z_j = 0 \Leftrightarrow \beta_j = 0$ into an optimization problem, such as:

1. Replace β_j by $z_j \beta_j$ for $j = 1, \dots, p$.
2. Set $|\beta_j|(1 - z_j) = 0$ for $j = 1, \dots, p$ or $\sum_{j=1}^p |\beta_j|(1 - z_j) = 0$.
3. Use a big- M constraint, $|\beta_j| \leq M_v z_j$ for $j = 1, \dots, p$ and for some fixed constant M_v large enough (such as $M_v \geq \max_j |\beta_j^*|$, β_j^* being the solution of the optimization problem). In the setup of experimental results for synthetic data sets, we explain how we can set a priori value of M_v .
4. Treat $z_j = 0 \Leftrightarrow \beta_j = 0$ as logical implications (also called indicator constraints or special ordered set SOS-1). Note that this kind of logical implication can be efficiently handled in a branch-and-bound procedure for MIO problems.

We now discuss and give a short overview of the advantages and drawbacks of each approach. The two first approaches involve nonlinear interaction terms between binary and continuous variables. Their interest lies in the possibility of obtaining interesting

continuous relaxations. The main advantage of the big- M method (approach 3) is that it brings only linear inequality constraints, but the value of the M term needs to be chosen carefully since it shows a great deal of practical influence on the solver performance. Logical implications (approach 4) have the advantage of avoiding these types of problems, as they do not rely on a separate constant value. However, they tend to have weaker relaxations, a condition which may lead to longer solve times in a model. In this paper we will use the third approach for our implementation since the presented discrete first order algorithm allows to obtain a good upper bound of M and since the brought linear inequality constraint do not have a significant influence on the computational time.

Outlier detection also involves the ℓ_0 norm function to count the number of outliers. As done above, this counting function can be represented by introducing n binary variables $t_i \in \{0, 1\}$ such as

$$\|\tau\|_0 = \sum_{i=1}^n t_i \quad \text{and} \quad t_i = 0 \Rightarrow \tau_i = 0, (x_i, y_i) \text{ is not an outlier.}$$

2.2 A MIO Formulation

Introducing binary variables for both variables and outliers with two big- M constraints, given appropriate parameters k_v, k_o, M_v and M_o , Problem (3) becomes for some $q \in \{1, 2\}$:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p, \tau \in \mathbb{R}^n, z \in \{0, 1\}^p, t \in \{0, 1\}^n} \quad & \frac{1}{q} \|X\beta + \tau - y\|_q^q \\ \text{s.t.} \quad & \sum_{j=1}^p z_j \leq k_v \quad \text{and} \quad |\beta_j| \leq z_j M_v, \quad j = 1, \dots, p \\ & \sum_{i=1}^n t_i \leq k_o \quad \text{and} \quad |\tau_i| \leq t_i M_o \quad i = 1, \dots, n. \end{aligned} \quad (6)$$

This problem turns out to be a mixed binary quadratic program when $q = 2$, it will be denoted by ℓ_0 -RR and it will be used in the rest of the paper. However, we will introduce other formulations that could also be efficient without using these formulations in the experiments.

2.3 Convergence to the Global Optimum

Figure (1) shows the influence of the SNR value on the speed of convergence. In fact, we consider a synthetic data set without adding outliers. When $k_o = 5\%$, the time needed to certify the optimality decreased from 120 seconds for $SNR = 0.5$ to 52 seconds for $SNR = 5$. In addition, after three hours of computation and when $k_o = 10\%$, the MIO-Gap decreased from 0.2 ($SNR = 0.5$) to 0.1 ($SNR = 5$).

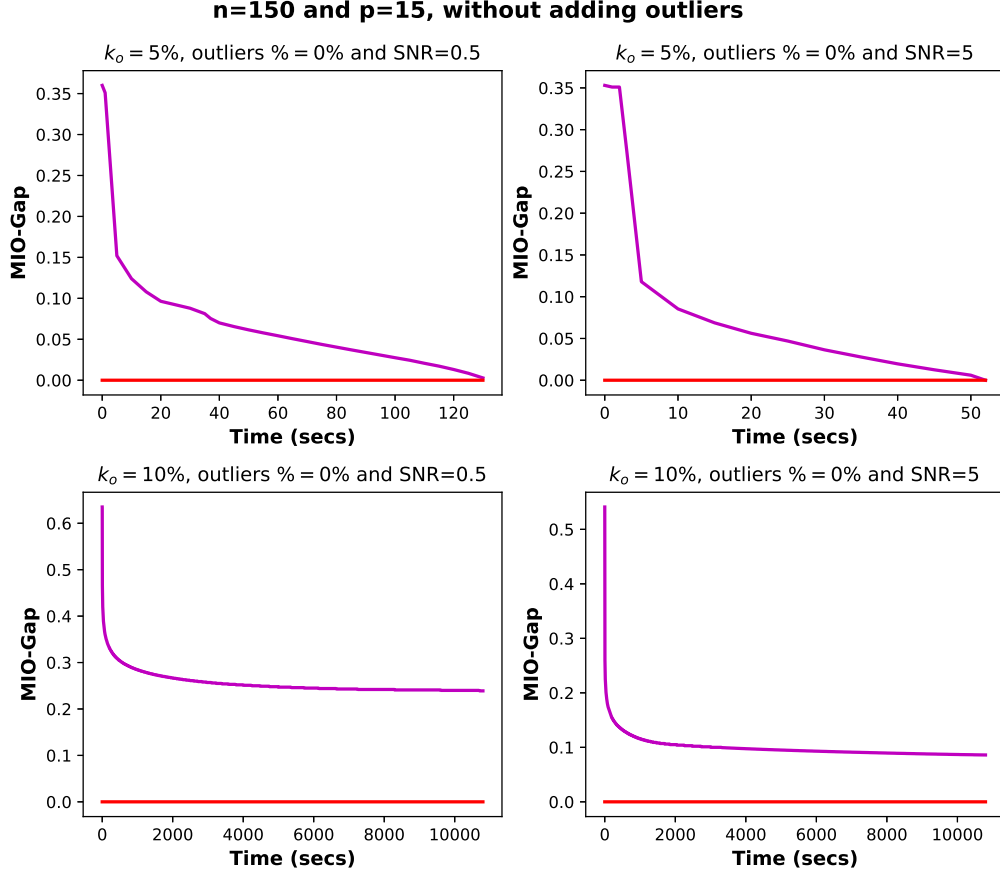


Figure 1: The typical evolution of the MIO formulation (6) for a synthetic dataset with $n = 150$, $p = 15$, $s = 5$. The top and the bottom panels show the evolution of the corresponding MIO gap with time. The red line is the $y = 0$ reference line.

In Figure (2) we shed light on the importance of estimating the true percentage of outliers in the data set (10% in our case). When we set k_o as the true percentage of outliers (right panel), the optimality was certified in about three minutes. But when the true percentage of outliers is underestimated ($k_o = 2.5\%$), the MIO-Gap was still about 0.2 even after 3 hours. Note that when we overestimate the percentage of outliers ($k_o = 15\%$ for example) we observe slow convergence as we did when underestimating it.

In summary, the convergence rate depends on many factors:

- the size of the data set: smaller data leads to faster convergence to optimality,
- the estimation of the parameters k_v and k_o : better estimation of the number of relevant features and of the percentage of outliers increases the speed of convergence to optimality,
- the noise in the data (SNR): more time is needed to certify optimality for lower SNR values.

n=500 and p=100, with 10% of outliers

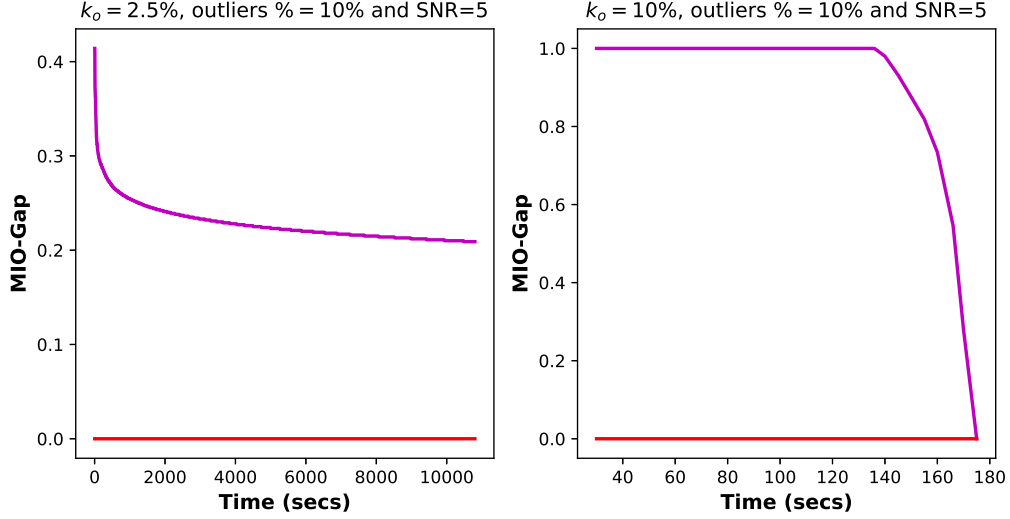


Figure 2: The typical evolution of the MIO formulation (6) for a synthetic dataset with $n = 500, p = 100, s = 5$. The left and the right panels show the evolution of the corresponding MIO gap with time. The red line is the $y = 0$ reference line.

3 Proximal Alternating Linearized Minimization Algorithm

In this section, an efficient alternate projected gradient algorithm providing a local solution to the optimization Problem (3) is introduced. This algorithm will be used as a warm-start procedure for the MIO solver as well as an optimization algorithm itself since it could provide high quality solutions in a short time. Before entering into the details of the alternate projected gradient algorithm, it is appropriate to introduce the problem of finding the projection of a vector $u \in \mathbb{R}^p$ onto the set of $k \leq p$ sparse vectors

$$\begin{aligned} \min_{v \in \mathbb{R}^p} \quad & \frac{1}{2} \|v - u\|^2 \\ \text{s.t.} \quad & \|v\|_0 \leq k. \end{aligned} \quad (7)$$

This problem is easy and its solution v^* is given by sorting on the absolute value of vector $|u|$, that is by a sequence of indices (j) such that $|u_{(1)}| \geq |u_{(2)}| \geq \dots |u_{(j)}| \geq \dots \geq |u_{(p)}|$. Using these indices, the projection $v^* = P_k(u)$ of u is the vector u itself with its smallest coefficients set to 0 that is

$$v^* = P_k(u) = \begin{cases} u_j & \text{if } j \in \{(1), \dots, (k)\} \\ 0 & \text{else.} \end{cases} \quad (8)$$

We propose to use this projection mechanism, on both β and τ , to get a solution to the initial Problem (3) at a low computing cost.

A possible way to achieve this goal consists of using the so-called block Gauss-Seidel iteration scheme on variables β and τ , also known as alternating minimization. To this end, a sequence $\{(\beta^\ell, \tau^\ell)\}_{\ell \in \mathbb{N}}$ is generated starting from some (β^0, τ^0) using the following scheme:

$$\begin{cases} \beta^{\ell+1} = \arg \min_{\beta \in \mathbb{R}^p} (\beta - \beta^\ell)^\top X^\top (X\beta^\ell + \tau^\ell - y) \\ \text{s.t. } \|\beta\|_0 \leq k_v \\ \|\beta - \beta^\ell\|^2 \leq d_v \end{cases} \quad \begin{cases} \tau^{\ell+1} = \arg \min_{\tau \in \mathbb{R}^n} (\tau - \tau^\ell)^\top (X\beta^{\ell+1} + \tau^\ell - y) \\ \text{s.t. } \|\tau\|_0 \leq k_o \\ \|\tau - \tau^\ell\|^2 \leq d_o. \end{cases} \quad (9)$$

$$\begin{cases} \frac{1}{2} \|X\beta + \tau^\ell - y\|^2 \leq \frac{1}{2} \|X\beta^\ell + \tau^\ell - y\|^2 + (\beta - \beta^\ell)^\top X^\top (X\beta^\ell + \tau^\ell - y) + \frac{1}{2\rho_v} \|\beta - \beta^\ell\|^2 \\ \frac{1}{2} \|X\beta^{\ell+1} + \tau - y\|^2 \leq \frac{1}{2} \|X\beta^{\ell+1} + \tau^\ell - y\|^2 + (\tau - \tau^\ell)^\top (X\beta^{\ell+1} + \tau^\ell - y) + \frac{1}{2\rho_o} \|\tau - \tau^\ell\|^2. \end{cases} \quad (10)$$

Where d_v and d_o are two given positive parameters that can be changed each step. The idea of the proximal method is, at each iteration, to minimize a regularized first-order approximation of the cost that can be interpreted as a local trust region mechanism (for details see for instance Parikh and Boyd, 2014). This surrogate loss is also a local upper bound of the targeted loss since, for well chosen ρ_v and ρ_o , the Lagrange multipliers associated with the trust region constraints

$$\begin{cases} \frac{1}{2} \|X\beta + \tau^\ell - y\|^2 \leq \frac{1}{2} \|X\beta^\ell + \tau^\ell - y\|^2 + (\beta - \beta^\ell)^\top X^\top (X\beta^\ell + \tau^\ell - y) + \frac{1}{2\rho_v} \|\beta - \beta^\ell\|^2 \\ \frac{1}{2} \|X\beta^{\ell+1} + \tau - y\|^2 \leq \frac{1}{2} \|X\beta^{\ell+1} + \tau^\ell - y\|^2 + (\tau - \tau^\ell)^\top (X\beta^{\ell+1} + \tau^\ell - y) + \frac{1}{2\rho_o} \|\tau - \tau^\ell\|^2. \end{cases} \quad (11)$$

For each iteration, this method introduced by Bolte, Sabach and Teboulle (2014) and called the proximal alternating linearized minimization (PALM) algorithm, consists of minimizing the upper bounds as follows:

$$\begin{cases} \beta^{\ell+1} = \arg \min_{\beta \in \mathbb{R}^p, \|\beta\|_0 \leq k_v} (\beta - \beta^\ell)^\top X^\top (X\beta^\ell + \tau^\ell - y) + \frac{1}{2\rho_v} \|\beta - \beta^\ell\|^2 \\ \tau^{\ell+1} = \arg \min_{\tau \in \mathbb{R}^n, \|\tau\|_0 \leq k_o} (\tau - \tau^\ell)^\top (X\beta^{\ell+1} + \tau^\ell - y) + \frac{1}{2\rho_o} \|\tau - \tau^\ell\|^2. \end{cases} \quad (12)$$

That is, after some algebra,

$$\begin{cases} \beta^{\ell+1} = \arg \min_{\beta \in \mathbb{R}^p, \|\beta\|_0 \leq k_v} \frac{1}{2} \|\beta - \beta^\ell + \rho_v X^\top (X\beta^\ell + \tau^\ell - y)\|^2 \\ \tau^{\ell+1} = \arg \min_{\tau \in \mathbb{R}^n, \|\tau\|_0 \leq k_o} \frac{1}{2} \|\tau - \tau^\ell + \rho_o (X\beta^{\ell+1} + \tau^\ell - y)\|^2. \end{cases} \quad (13)$$

These two minimization problems are of the same kind as Problem (7) and thus the sequence can be generated by using two ℓ_0 projected gradient, that is:

$$\begin{cases} \beta^{\ell+1} = P_{k_v}(\beta^\ell - \rho_v X^\top (X\beta^\ell + \tau^\ell - y)) \\ \tau^{\ell+1} = P_{k_o}(\tau^\ell - \rho_o (X\beta^{\ell+1} + \tau^\ell - y)). \end{cases} \quad (14)$$

Algorithm 1 presents the pseudo code of the PALM algorithm.

Algorithm 1: Proximal alternating linearized minimization (PALM) Bolte et al. (2014)

Data: X, y initialization $\beta, \tau = 0$

Result: β, τ

set $\rho_v \leq \frac{1}{\sigma_M^2}$ and $\rho_o \leq 1$

while it has not converged ($\|\beta_{n+1} - \beta_n\|_2 > 10^{-6}$) **do**

$d \leftarrow \beta - \rho_v X^\top (X\beta + \tau - y)$

variable selection

$\beta \leftarrow P_{k_v}(d)$

$\delta \leftarrow \tau - \rho_o (X\beta + \tau - y)$

eliminating outliers

$\tau \leftarrow P_{k_o}(\delta)$

This algorithm converges towards a local minima of Problem (3) since it fulfills the assumptions needed for Theorem 3.1 in Bolte et al. (2014). Indeed, if we consider $G(\beta, \tau) = \frac{1}{2} \|X\beta + \tau - y\|_2^2$, PALM converges if the partial gradients $G_\beta(\beta) = X^\top (X\beta + \tau - y)$ and $G_\tau(\tau) = (X\beta + \tau - y)$ are globally Lipschitz with modules L_1 and L_2 respectively. It could be easily shown that $G_\beta(\beta)$ and $G_\tau(\tau)$ are $\frac{1}{\sigma_M^2}$ and 1 Lipschitz respectively, σ_M being the largest singular value of X . Thus the step sizes could be chosen such that $\rho_v \leq \frac{1}{\sigma_M^2}$ and $\rho_o \leq 1$ as proved in Bolte et al. (2014).

4 Results for Synthetic Data Sets

In this section we show the empirical performance of the MIO approach.

4.1 Setup

In Hastie et al. (2017), a follow-up paper to Bertsimas et al. (2015), the authors provide a synthetic setup considering a wide range of SNR values. We use it here to compare the best subset selection (Formulation (6) with $k_o = 0$), the lasso, PALM, the ℓ_0 robust regression - ℓ_0 RR and the ℓ_1 robust regression - ℓ_1 RR. The same notations as Hastie et al. (2017) were used, namely n, p (problem dimensions), s (sparsity level), beta-type (pattern of sparsity), ρ (predictor auto-correlation level) which controls correlations between predictor variables, and ν (SNR level).

- We define coefficients $\beta_0 \in \mathbb{R}^p$ according to s and the beta-type, as described below.
- We draw the rows of the matrix $X \in \mathbb{R}^{n \times p}$ from $N_p(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ has entry (i, j) equal to $\rho^{|i-j|}$, and $\rho = 0.35$.

- We draw the vector $y \in \mathbb{R}^n$ from $N_n(X\beta_0, \sigma^2 I)$, with σ^2 defined to meet the desired SNR level, i.e., $\sigma^2 = \beta_0^T \Sigma \beta_0 / \nu$.
- We use 5-fold cross-validation and the tuning was performed by minimizing prediction error on the test set.
- To assess the influence of outliers, 5% of outliers were added to the data set by following a normal $N(50, \sigma)$ instead of $N(0, \sigma)$.
- We considered two configurations: the low setting with $n = 150, p = 15$, and the medium setting $n = 500, p = 100$. For each configuration, we also considered two settings: the first one with outliers generated as mentioned above, and the second one without adding outliers.
- The lasso was tuned over 100 values of λ (as it is in `glmnet`).
- In order to determine the values of k_v, M_v, k_o and M_o , we run the PALM algorithm for k_v ranging from 1 to p and for k_o ranging from 0 to 10% with a step size of 2.5%. Then, we choose the solution with the minimal error $\|X_{test}\beta_{palm} - y_{test}\|_2^2$.
- $M_v = (1 + \alpha)\|\beta_{palm}\|_\infty, M_o = (1 + \alpha)\|\tau_{palm}\|_\infty$ with $\alpha = 0.1$, k_v and k_o are set as the number of nonzero elements in the solutions β_{palm} and τ_{palm} respectively.
- The ℓ_1 robust regression (ℓ_1 RR) algorithm was tuned over five values of λ from zero to $1.5\|\beta_{lasso}\|_\infty$ where β_{lasso} is the solution obtained by the lasso method, and over fifty one values of γ from 0 to 5000 with a step size of 100 for the low dimensional case, and from 0 to 10000 with a step size of 200 for the medium dimensional case.
- We run the best subset selection, the lasso, PALM, the ℓ_0 robust regression (ℓ_0 RR) the ℓ_1 robust regression (ℓ_1 RR) using a 5-fold cross-validation. The tuning was performed by minimizing the error on the test set.
- We repeat 10 times for the low dimensional setting and 5 times for the medium dimensional setting and average the results.

Coefficients: We considered three settings for the coefficients $\beta_0 \in \mathbb{R}^p$ as in Hastie et al. (2017):

- beta-type 1: $\beta_0 = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, \underbrace{0, \dots, 0}_{p-10 \text{ times}})$;
- beta-type 2: β_0 has its first 5 components equal to 1, and the rest equal to 0;
- beta-type 5: β_0 has its first 5 components equal to 1, and the rest decaying exponentially to 0, specifically, $\beta_{0i} = 0.5^{i-s}$, for $i = s + 1 \dots p$, where $s = 5$;

Following Bertsimas et al. (2015); Hastie et al. (2017), we use, as an accuracy metric, the relative risk (R.R) defined by:

$$\text{R.R}(\hat{\beta}) = \frac{\mathbb{E}(x_0^T \hat{\beta} - x_0^T \beta_0)^2}{\mathbb{E}(x_0^T \beta_0)^2} = \frac{(\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0)}{\beta_0^T \Sigma \beta_0},$$

The best score is 0 (when $\hat{\beta} = \beta_0$) and the null score is 1, obtained when $\hat{\beta} = 0$.

We also use the proportion of variance explained (PVE) defined by:

$$\text{PVE}(\hat{\beta}) = 1 - \frac{\mathbb{E}(y_0 - x_0^T \hat{\beta})^2}{\text{Var}(y_0)} = 1 - \frac{(\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0) + \sigma^2}{\beta_0^T \Sigma \beta_0 + \sigma^2}.$$

The maximum value for the PVE, also called the perfect score, is $\text{SNR}/(\text{SNR}+1)$ (see Hastie et al. (2017) for details).

4.2 Computational Costs

For the lasso, we used the **Matlab** “lasso” function with 100 values of λ as implemented in **glmnet**. The solution is delivered in a very short time. For the best subset selection problem, we implemented the method using the MIO Formulation (6) with $k_o = 0$, used PALM to compute a warm start and then call Gurobi through its Matlab interface. We used a time limit of 3 minutes for Gurobi to optimize the best subset selection problem for both low and medium dimensional case. The same procedure is followed for the ℓ_0 robust regression problem but with a time limit increased to 10 minutes for the medium dimensional setting.

For the ℓ_1 robust regression, we obtained $5 \times 51 = 255$ (5 values of λ and 51 values of γ) solutions for each test. The time needed to obtain each solution depends on the size of the dataset, but it varies from 0.16 second to about 1 second.

We can conclude that for low dimensional setting, we faced around 15 hours of computation (10 repetitions), and more than 45 hours for the medium dimensional setting (5 repetitions) for each type of β . Using only one cross-validation loop would decrease significantly the computational time of the experiments. We note that the computations were carried on in a windows 10 64-bit server - Intel(R) Core(TM) i7-4700MQ CPU @ 2.40 GHz and 8 GB of Ram. So using a more powerful machine would help to decrease the computational cost.

4.3 Results

Figures (3)-(8) plot the relative risk (left panel) and the proportion of variance explained (right panel) as functions of signal-to-noise ratio (SNR). The results can be divided into two main categories:

4.3.1 No Outliers

In this case, no outliers were added to the synthetic data sets generated as mentioned before. Figures (3), (4), (5), (6), (7) and (8) show that for small SNR values, the ℓ_1 methods (lasso and ℓ_1 RR) have the lead on the other methods (best subset selection, PALM and ℓ_0 RR). While for high SNR values the ℓ_0 approaches outperform the ℓ_1

approaches even though all the methods perform quite similarly for high SNR values. These results shed the light on the capability of the MIO approach to perform well when no outliers exist in the data set.

4.3.2 Presence of Outliers

In this case, Figures (3), (4), (5), (6), (7) and (8) show that PALM, ℓ_0 RR and ℓ_1 RR outperform the best subset selection and the lasso, which is not surprising since the last two methods are not robust to outliers. The obtained results ensure that adding the variable τ helped to improve the performance of the estimators and guaranteed obtaining robust methods. In addition, for $SNR < 0.25$ the ℓ_1 RR performs, in general, better than PALM and the ℓ_0 RR. But for higher SNR values, there is no clear winner. An important caveat to emphasize up front is that the Gurobi MIO algorithm for ℓ_0 RR was given only

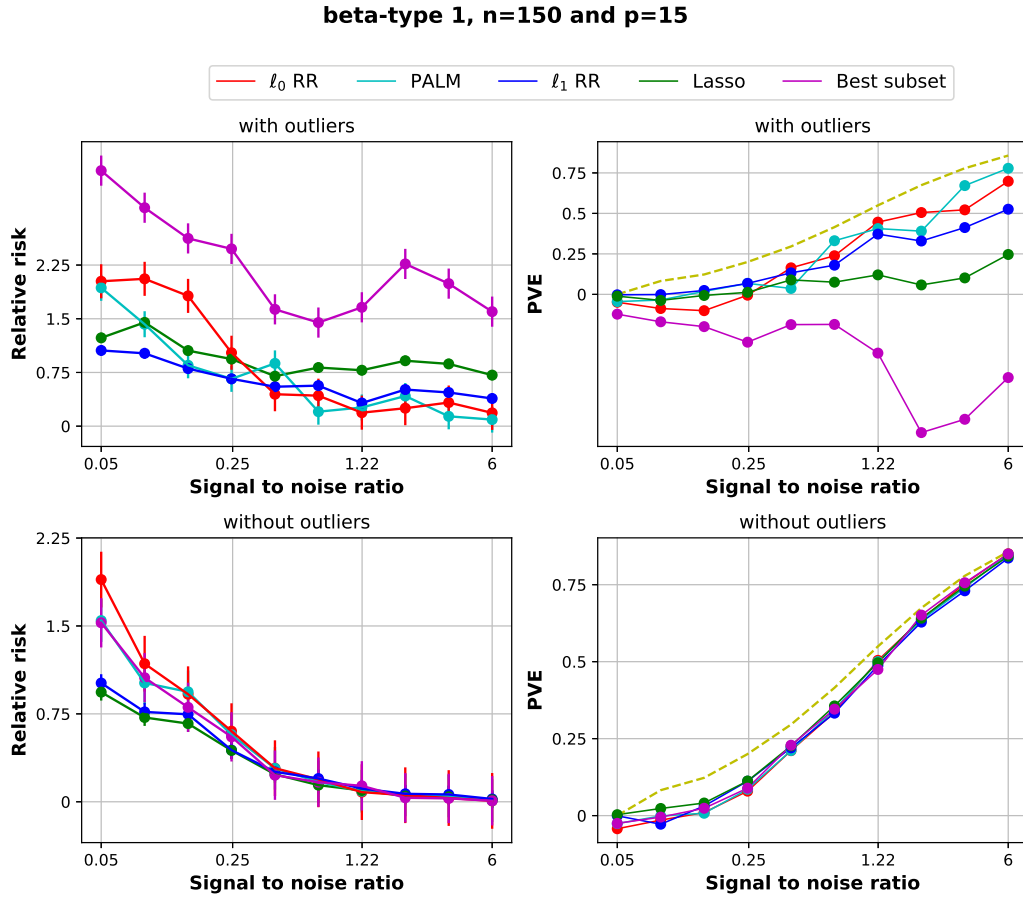


Figure 3: Relative risk (left panel) and proportion of variance explained (right panel) functions of SNR, for beta-type 1 in the setting with $n = 150$, $p = 15$, and $s = 5$ with and without outliers (top panel and bottom panel respectively).

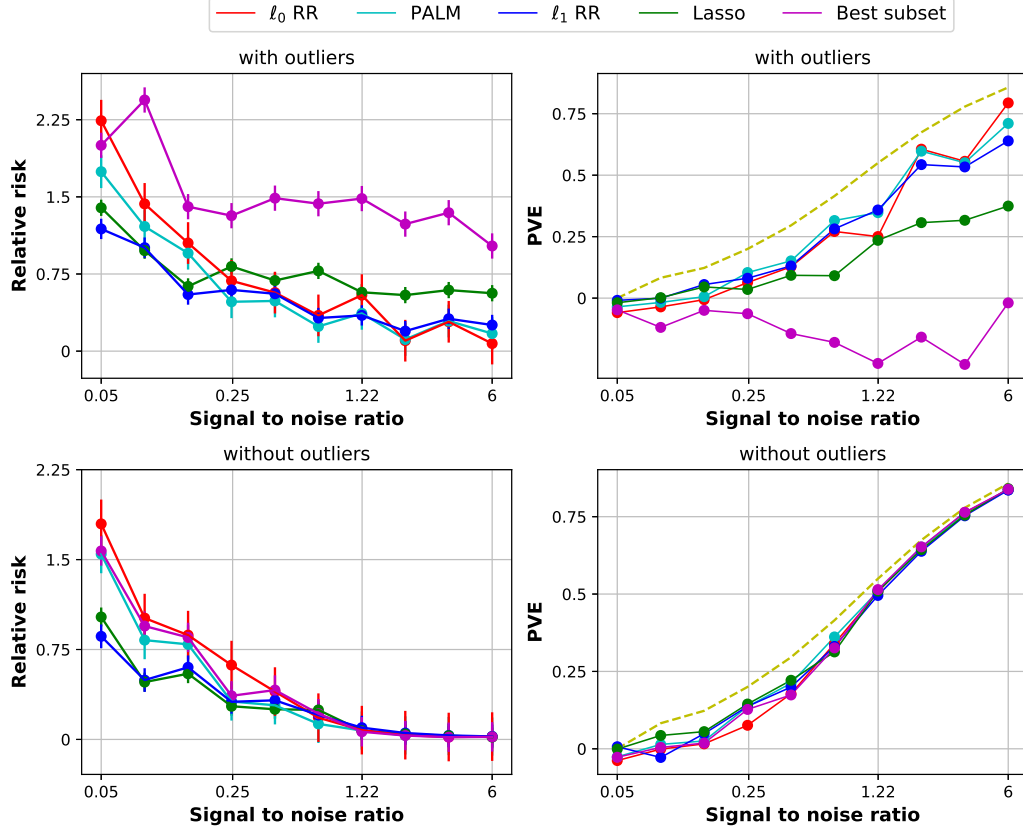
beta-type 2, n=150 and p=15

Figure 4: Relative risk (left panel) and proportion of variance explained (right panel) functions of SNR, for beta-type 2 in the setting with $n = 150$, $p = 15$, and $s = 5$ with and without outliers (top panel and bottom panel respectively).

10 minutes per problem, which may have caused the ℓ_0 RR to underperform, and that the performance of the MIO algorithm depends on the parameters tuned using PALM.

4.4 Detection Rate for the Feature Selection and Outlier Detection Tasks

To determine whether the ℓ_0 robust regression approach can detect the outliers and select the right features, we generated two low-dimensional and two medium-dimensional data sets using the β type-2, with SNR values 0.5 and 5. We added 5% of outliers in the response vector (as in the setup of the synthetic data sets). k_v and k_o were set as the true sparsity level of β and as the percentage of outliers (5%). In all cases, the detection rate of both outliers and features was 100%, noting that no cross-validation was performed.

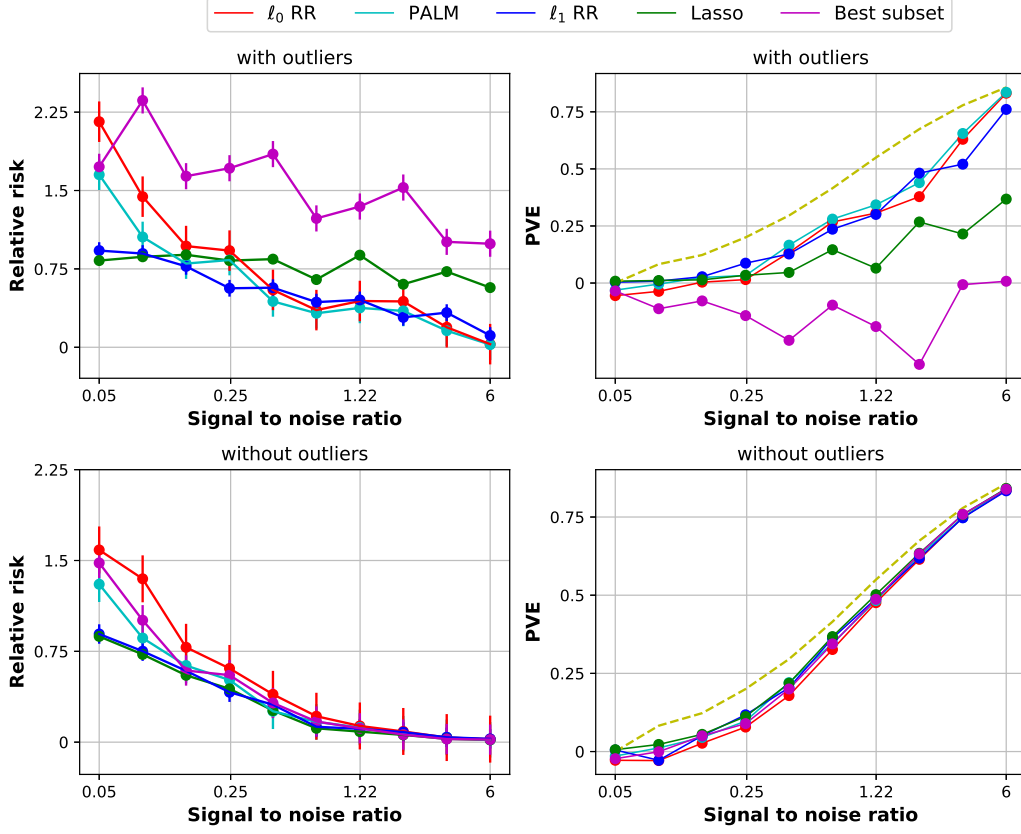
beta-type 5, $n=150$ and $p=15$ 

Figure 5: Relative risk (left panel) and proportion of variance explained (right panel) functions of SNR, for beta-type 5 in the setting with $n = 150$, $p = 15$, and $s = 5$ with and without outliers (top panel and bottom panel respectively).

In the experiments performed on both real and data sets, we used PALM to tune the parameters k_v and k_o . Thus the performance of the MIO approach depends on PALM. To this end, each data set was split into two parts: the training set (70%) and the testing set (30%). We added 5% of outliers in the training set's response vector. PALM was performed for $k_v \in [1, \dots, p]$ and $\frac{k_o}{n} \in [0, 0.025, 0.05, 0.075, 0.1]$. PALM failed to estimate the true sparsity level and the true percentage of outliers as seen in Figures (9) and (10). This leads the PALM-MIO approach to fail at detecting the percentage of outliers and selecting the correct number of relevant features, even though all the true outliers were considered as outliers by this approach.

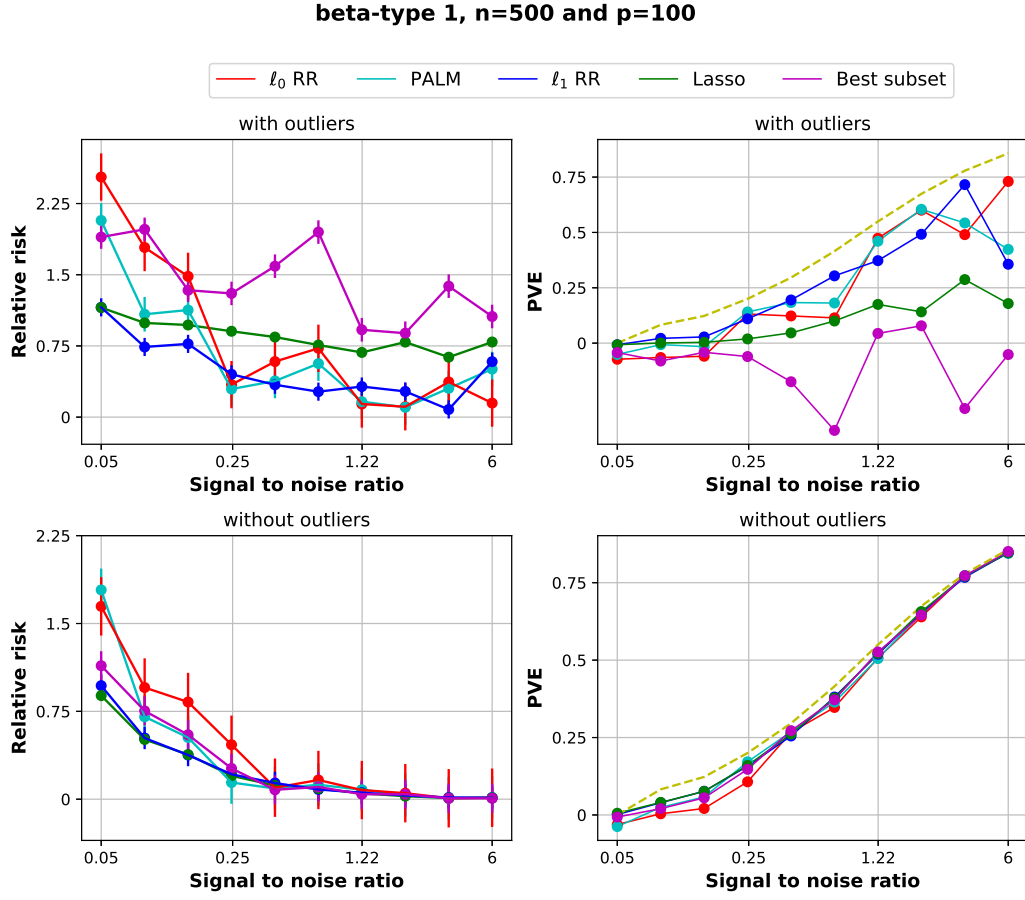


Figure 6: Relative risk (left panel) and proportion of variance explained (right panel) functions of SNR, for beta-type 1 in the setting with $n = 500$, $p = 100$, and $s = 5$ with and without outliers (top panel and bottom panel respectively).

5 Real Data Sets

The performances of all methods have been compared on real data sets. To this end we have used 7 data sets presented in Table 1. The different methods have been compared on all these data sets according to the following setup:

- The response vector y and the columns of the matrix X have been standardized to have zero mean and unit standard deviation;
- Two 5-fold cross-validation loops have been implemented. The inner one has been used to give a relevant choice for the hyper-parameters. The outer one has been used to estimate the average mean squared error MSE;

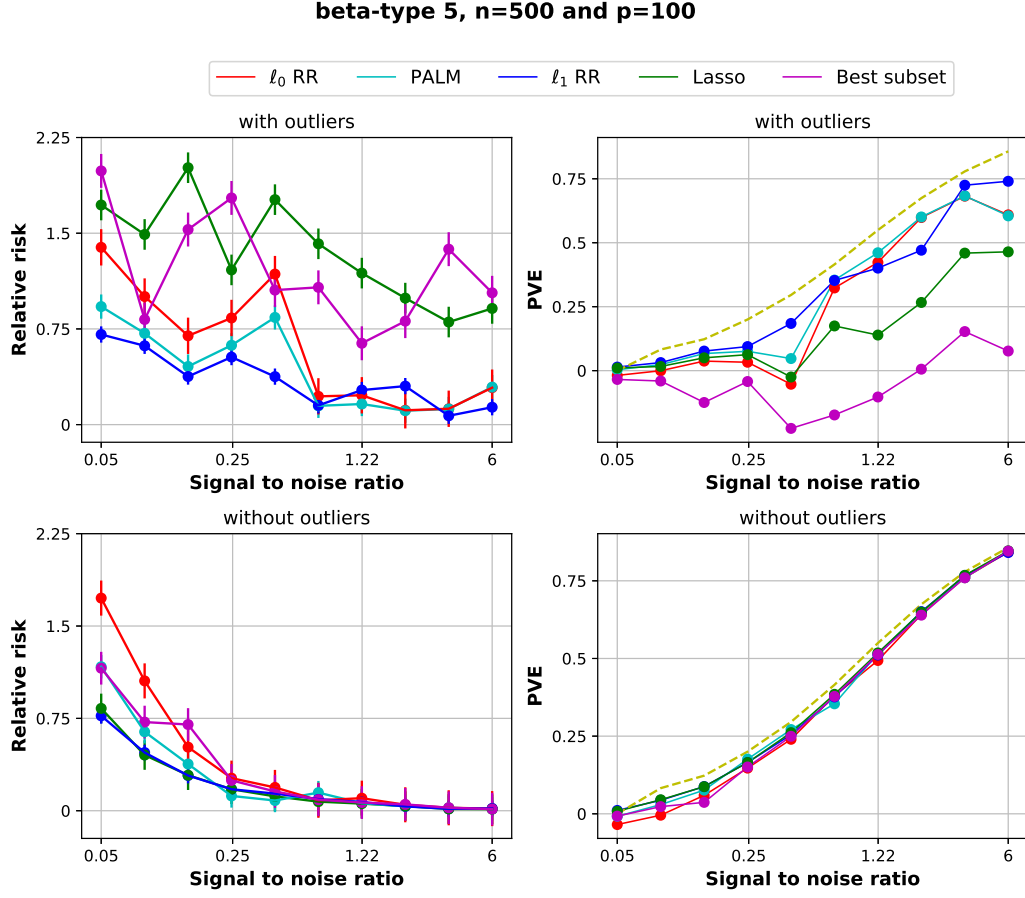


Figure 7: Relative risk (left panel) and proportion of variance explained (right panel) functions of SNR, for beta-type 2 in the setting with $n = 500$, $p = 100$, and $s = 5$ with and without outliers (top panel and bottom panel respectively).

- As for synthetic data sets, we run PALM for k_v ranging from 1 to p , and k_o ranging from 0 to 10% with a step size of 2.5%, and pick the solution with smallest cross-validation error. This obtained solution is used to set the values of M_v and M_o and as a warm start for the ℓ_0 robust regression algorithm as well;
- The hyper-parameter λ of the lasso was tuned over 100 values as per the default in `glmnet`;
- The ℓ_1 robust regression algorithm was tuned over 5 values of λ (as for the synthetic data sets) and over 40 values of γ varying from 0 to 2000 with a step size of 50. We remarked that, for the normalized and standardized data set considered, it's enough to bound $\|\tau\|_1$ by 2000;
- Outliers were generated by replacing 5% of the response vector values y_i by $y_i + 2(\max(y) - \min(y))$ that is a constant value set to the range of the response variable in the training set;

beta-type 2, n=500 and p=100

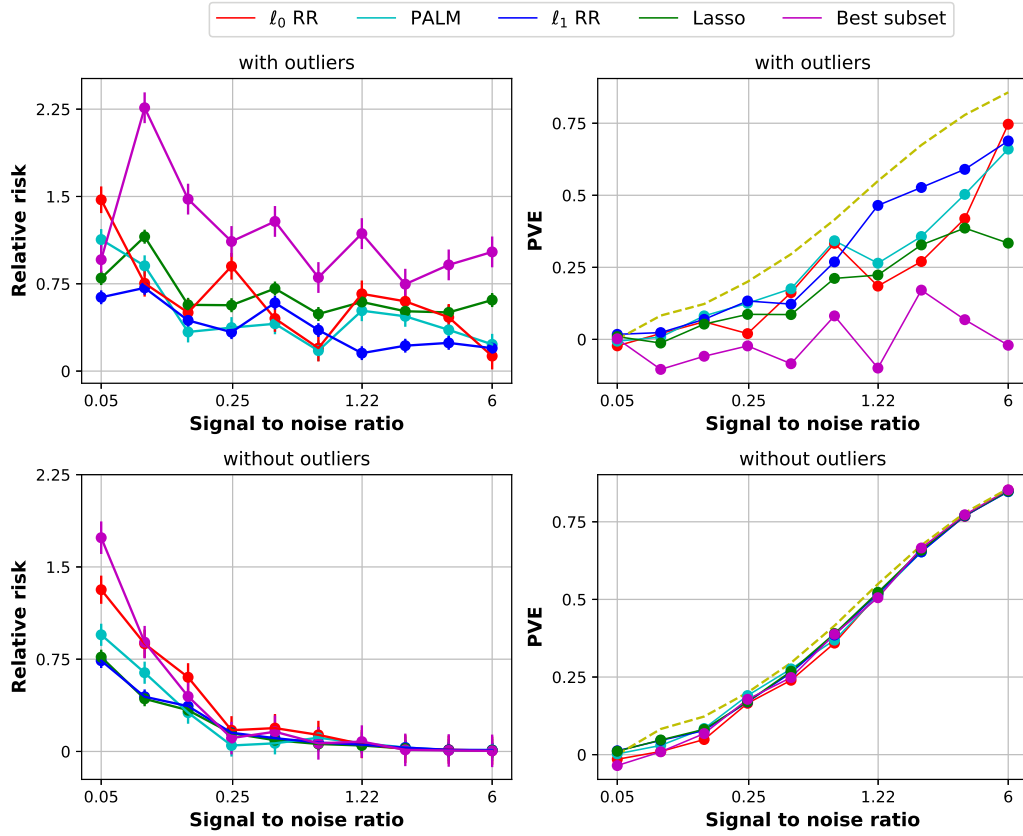


Figure 8: Relative risk (left panel) and proportion of variance explained (right panel) functions of SNR, for beta-type 5 in the setting with $n = 500$, $p = 100$, and $s = 5$ with and without outliers (top panel and bottom panel respectively).

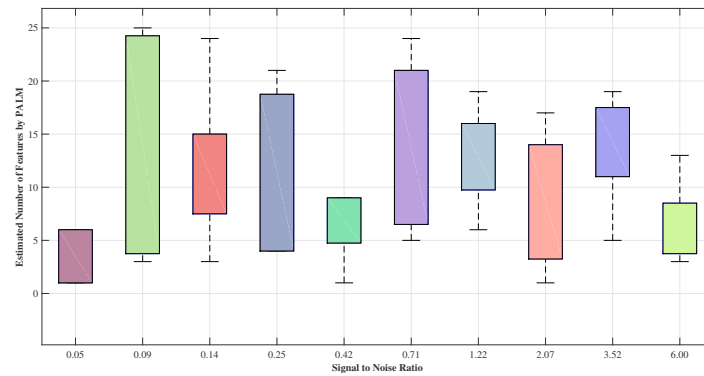


Figure 9: Summary of used datasets.

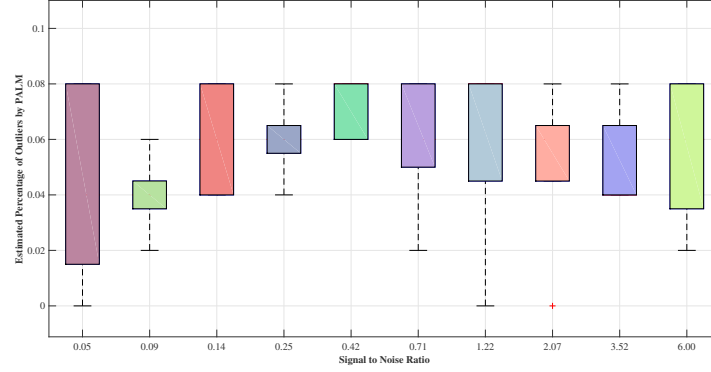


Figure 10: Percentage of outliers over estimation by PALM. The percentage of outliers in the data set is 5%.

Each experiment is repeated 3 times. Tables 2 and 3 report the average of the results and the standard deviation in parentheses for the raw data.

Table 1: Periods and sites extracted from clear archaeological contexts with radiocarbon determinations.

Name of the dataset	number of instances n	number of attributes p	Origin
Body Fat	252	15	lib.stat.cmu.edu
Concrete Compressive Strength	1030	9	UCI
Concrete Slump Test	103	10	UCI
Real Estate Valuation	414	7	UCI
Diabetes	442	10	stat.ncsu.edu
Boston Housing	489	3	Web ¹
Auto Mpg	398	8	UCI

Table 2: Cross-validation MSE rates (standard deviations) of the best subset, lasso, PALM, ℓ_0 robust regression (ℓ_0 RR) and ℓ_1 robust regression (ℓ_1 RR) on 7 real datasets.

	Best subset	Lasso	Palm	ℓ_0 RR	ℓ_1 RR
Body Fat	2.2797 ($7.2e^{-5}$)	4.2644 ($1.5e^{-4}$)	2.5958 ($5.2e^{-5}$)	2.6270 ($4.77e^{-5}$)	4.5008 ($6.2e^{-5}$)
Concrete Compressive Strength	0.3588 (0.018)	0.3602 (0.019)	0.3692 ($4.2e^{-4}$)	0.3693 ($3.5e^{-4}$)	0.3603 (0.015)
Slump Test	0.0880 (0.008)	0.0863 (0.012)	0.0864 (0.011)	0.0880 (0.008)	0.0869 (0.010)
Real Estate Valuation	0.2994 (0.024)	0.2924 (0.036)	0.3010 (0.026)	0.2992 (0.026)	0.2950 (0.033)
Diabetes	0.3917 (0.037)	0.3914 (0.038)	0.3889 (0.028)	0.3888 (0.038)	0.3952 (0.039)
Boston Housing	0.2460 (0.007)	0.2460 (0.007)	0.2446 (0.008)	0.2440 (0.009)	0.2448 (0.006)
Auto Mpg	0.1469 (0.002)	0.1458 (0.005)	0.1523 (0.007)	0.1516 (0.007)	0.1478 (0.008)

An important caveat to emphasize upfront is that the ℓ_0 robust regression algorithm was given 10 minutes time limit per problem instance per subset size. This practical restriction may have caused this algorithm to underperform in some cases. For the best subset selection problem, the time limit was set to 2 minutes. We note that the optimality was certified for almost every case in less than two minutes. In the absence of outliers, results in Table 2 show that there is no clear winner. It is remarkable that all methods

performed quite similarly, with a little advantage of using the lasso. In the presence of outliers, results in Table 3 show the dominance of the robust regression algorithms used over the best subset selection and the lasso. The ℓ_0 robust regression performed better than the other methods.

Table 3: Cross-validation MSE rates (standard deviations) of the best subset, lasso, PALM, ℓ_0 robust regression (ℓ_0 RR) and ℓ_1 robust regression (ℓ_1 RR) on 7 real datasets corrupted by 5% of outliers in the initial response vector y .

	Best subset	Lasso	Palm	ℓ_0 RR	ℓ_1 RR
Body Fat	0.3923 (0.023)	0.4039 (0.034)	0.3679 (0.024)	0.3764 (0.009)	0.3882 (0.023)
Concrete compressive strength	0.5891 (0.063)	0.5877 (0.059)	0.5843 (0.070)	0.5842 (0.071)	0.5857 (0.755)
Slump test	0.2749 (0.186)	0.2463 (0.128)	0.1110 (0.022)	0.0958 (0.012)	0.1039 (0.018)
Real estate valuation	0.6581 (0.131)	0.6680 (0.146)	0.6587 (0.137)	0.6580 (0.138)	0.6688 (0.147)
Diabetes	0.5087 (0.015)	0.5002 (0.011)	0.5012 (0.009)	0.5009 (0.011)	0.4923 (0.014)
Boston housing	0.5408 (0.240)	0.5293 (0.231)	0.5425 (0.241)	0.5441 (0.241)	0.5235 (0.225)
Auto mpg	0.5498 (0.139)	0.5596 (0.128)	0.5406 (0.160)	0.5406 (0.160)	0.5370 (0.163)

6 Conclusion

In this paper we propose a method for linear regression which solves the underlying optimization problem that handles both variable selection and outlier detection. We formulate the problem as a mixed-integer optimization problem and present a fast alternating minimization algorithm to find local minima. Furthermore, we present an empirical comparison between this method and its ℓ_1 relaxation on both synthetic and real data. We have found that neither the ℓ_0 norm problem nor its ℓ_1 relaxation dominates the other. Our recommendation is to use the ℓ_0 norm problem for large SNR while ℓ_1 relaxation is preferred when SNR is small. While the ℓ_0 approach is considered to be intractable, especially, for high dimensional regimes, one can propose to use screening rules helping in accelerating the solvers. Moreover, we have shown that if the true number of features and percentage of outliers are well estimated, the speed of convergence to the global minimum decreases significantly. Dealing with data sets of high dimensionality is the main limitation of the proposed MIO approach because of the high computational cost. However, we suggest to use the PALM algorithm in the high-dimensional case since it provides high quality solutions in a short time.

References

- Alfons, A., Croux, C. and Gelper, S. et al. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7, 226–248.
- Bertsimas, D., King, A. and Mazumder, R. (2015). Best subset selection via a modern optimization lens. *Annals of Statistics*, 47, 2324–2354.
- Bolte, J., Sabach, S. and Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146, 459–494.
- Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I. and Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, 30, 891–927.
- Chen, Y., Caramanis, C. and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pp. 774–782.
- Dalalyan, A. S. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1 -penalized huber’s m -estimator. *arXiv preprint arXiv:1904.06288*.
- Giloni, A. and Padberg, M. (2002). Least trimmed squares regression, least median squares regression, and mathematical programming. *Mathematical and Computer Modelling*, 35, 1043–1060.
- Hastie, T., Tibshirani, R. and Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.
- Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22, 85–126.
- Miller, A. (2002). *Subset selection in regression*. CRC Press.
- Miyashiro, R. and Takano, Y. (2015). Subset selection by mallows: A mixed integer programming approach. *Expert Systems with Applications*, 42, 325–331.
- Nguyen, N. H. and Tran, T. D. (2013). Robust lasso with missing and grossly corrupted observations. *IEEE transactions on information theory*, 59, 2036–2058.
- Öllerer, V., Alfons, A. and Croux, C. (2016). The shooting s -estimator for robust regression. *Computational Statistics*, 31, 829–844.
- Parikh, N. and Boyd, S. P. (2014). Proximal algorithms. *Foundations and Trends in optimization*, 1, 127–239.
- Rousseeuw, P. J. and Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, e1236.
- Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, Volume 589. John Wiley & Sons.
- She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106, 626–639.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.
- Tibshirani, R., Wainwright, M. and Hastie, T. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. Chapman and Hall/CRC.
- Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25, 347–355.
- Yang, M., Xu, L., White, M., Schuurmans, D. and Yu, Y.-I. (2010). Relaxed clipping: A global training method for robust regression and classification. In *Advances in neural information processing systems*, pp. 2532–2540.

The unilateral spatial autoregressive process for the regular lattice two-dimensional spatial discrete data

Azmi Chutoo¹, Dimitris Karlis², Naushad Mamode Khan¹
and Vandna Jowaheer³

Abstract

This paper proposes a generalized framework to analyze spatial count data under a unilateral regular lattice structure based on thinning type models. We start from the simple spatial integer-valued auto-regressive model of order 1. We extend this model in certain directions. First, we consider various distributions as choices for the innovation distribution to allow for additional overdispersion. Second, we allow for use of covariate information, leading to a non-stationary model. Finally, we derive and use other models related to this simple one by considering simplification on the existing model. Inference is based on conditional maximum likelihood approach. We provide simulation results under different scenarios to understand the behaviour of the conditional maximum likelihood. A real data application is also provided. Remarks on how the results extend to other families of models are also given.

MSC: 62H11, 62M30.

Keywords: Unilateral, Spatial, Regular, Lattice, Thinning.

1 Introduction

Problems with spatial count data occur in several disciplines. For example, consider the Human West Nile virus counts spread (Tevie, Bohara and Valdez, 2014), the infant death syndrome for the counties in North Carolina (Cressie and Chan, 1989), the number of vehicle burglary incidents in counties of Texas (Chun, 2014) and most recently, the cases and/or deaths of COVID-19 outbreak.

Spatial data usually viewed as an aggregation or average of the events of interest emanates from a lattice structure. There are two main broad ways of representing the spatial observations (Cressie, 1993). The first, and most common way, is when the ob-

¹ Department of Economics and Statistics, University of Mauritius, Mauritius.

² Department of Statistics, Athens University of Economics and Business, Greece.

³ Department of Mathematics, University of Mauritius, Mauritius.

Received: December 2020

Accepted: April 2021

servation is in the form of a single indexed variable obtained from an areal unit k within some defined boundaries and in which case the spatial event is denoted as Y_k , for the k th areal unit. In the second way, the spatial count observation is indexed in two dimensional forms, in terms of the location or site coordinate (or also termed as the latitude and longitude position (i, j)), denoted as Y_{ij} , situated over a regular or irregular grid. Under both representations, the spatial count observation is supplemented by a neighbourhood structure, defined by terms of areal units or sites within the lattice structure. The second form of representation is useful since the two-dimensional representation considers all border cells in the region of interest (see, e.g. Tjøstheim, 1978a; Basu and Reinsel, 1993). Details on these representations can be found in Cressie (1993, Chap 6).

In the analysis of spatial data, it is important to investigate the spatial dependence between observations from the different neighbouring areal units or sites. Next, such analysis can also shed light on the possible factors or effects influencing the spatial observations and these can include variables such as the distance metric, elevation, slope, rock type and land use fault types (see, e.g. Tobler, 1969).

Several models have been studied in the literature to analyze spatial processes. The majority of the literature treats spatial continuous or discrete processes involving areal units, while only few papers consider the spatial data with a coordinate system, especially for the discrete case. Besides, such spatial observations are seen mostly in the agricultural, disease mapping, environmental and in the field of criminology. Specifically, in agriculture, the plantation field is usually split into small areas or say, square grids or cells with location (i, j) and wherein each cell, the investigator is interested on the number of plants cultivated subject to factors influencing its cultivation (see Kruijer et al. (2007) and references therein) and along with how the plants in the different neighbouring cells impact on the harvest in the (i, j) th position. Similarly in epidemiology, researchers are often concerned on the factors influencing the number of infected or death cases as a result of an outbreak of a virus in a region and how this is affecting the number of cases in the neighbouring regions. Such data has been treated in Cressie and Chan (1989), Wakefield (2007) while some more examples can be found in Lawson et al. (1999) and Lawson and Williams (2001). Moreover, in environmental field, the occurrences of road traffic accidents at different segments also illustrate spatial data analysis. In fact, in a hotspot analysis conducted in Barcelona, it was shown, via the local Moran statistics, that road accidents are concentrated in close neighbouring areas that have a complex road network systems with large roundabouts (Alvarez, 2020). Some other related research include the works by Valverde and Jovanis (2008) and Satria and Castro (2016). Last but not least, Mburu and Bakillah (2016) reported on the number of vehicle burglary incidents in small neighbouring regions of London which were highly spatially autocorrelated. Their study also revealed several influential factors such as unemployment and crimes in these areas of London.

Unsurprisingly, there is influx of models for the areal-type spatial data that include mainly the class of conditional autoregressive (CAR) models (Besag, 1974) and its extensions to Intrinsic CAR (ICAR) (Besag and Kooperberg, 1995), the Besag-Yorke-

Mollie (BYM) (Besag, Mollié and York, 1991) models and among several other extended CAR-based models; a review can be found in Obaromi (2019).

In fact, for the regular lattice data of discrete nature that are represented in terms of the site coordinates, the only works appearing in the literature so far use the spatial integer-valued auto-regressive model of order 1 (SINAR(1)) by Ghodsi, Shitan and Bakouch (2012) and Ghodsi (2015). The model was constructed by introducing dependence between the observation of interest with its unilateral spatial neighbouring observations via the binomial thinning operator defined by Steutel and Harn (1979). The structure is similar to the observation-driven integer-valued time series models defined in McKenzie (1986). Properties of the model including asymptotic properties of the CML estimators were thereon established and the spatial process was proven to be stationary and ergodic.

In the present paper we extend the model in certain directions. Firstly, we introduce different distributions for the innovations to enlarge the model and allow for larger variance, usually observed in spatial data due to clustering effects. Secondly, we allow for further spatial information to be used in the form of covariates that affect the model, leading to a non-stationary model. For this new model we discuss inference based on the CML. Moreover we discuss and apply some models related to the basic one that are parsimonious and easier to interpret, while they allow for easier extension to a broader family of models. Throughout the paper, some computational issues arising are also discussed.

The remaining of the paper proceeds as follows: The basic model and its extensions are described in Section 2. Simulations to further support the approach are provided in Section 3. A real data application related to the new models is provided in Section 4. Extensions of the current model and concluding remarks can be found in Section 5.

2 Generalised SINAR(1) model (GSINAR(1))

We consider spatial processes defined on a regular rectangular grid in two dimensions with sites labelled (i, j) , with an associated random variable Y_{ij} defined at each site. Examples of such phenomena include data collected on a regular grid of size $n_1 \times n_2$ from satellites and from agricultural field trials. The unilateral model (see, e.g. Tjøstheim, 1978b, 1983; Basu and Reinsel, 1993) defines the neighbouring sites that provide information for the site (i, j) , namely we denote as S_{ij} the set of indices (k, ℓ) of sites that are considered as neighbours of the site (i, j) and we define this as

$$S_{ij} = \{(k, \ell) \in \mathbb{Z}^2 : k \leq i, \ell \leq j\} - \{(i, j)\}.$$

Tjøstheim (1983) described the model of order (p_1, p_2) for continuous data as:

$$Y_{ij} = \sum_{k=0}^{p_1} \sum_{\ell=0}^{p_2} \phi_{k\ell} Y_{i-k, j-\ell} + \epsilon_{ij},$$

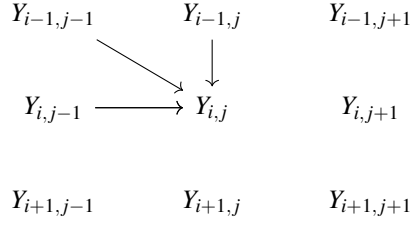


Figure 1: A diagram representing the unilateral model. The arrows indicate which site influence the site under consideration. Only sites located at the left and upper from the site into consideration influence the site.

where $\phi_{00} = 0$ and the errors ϵ_{ij} follow a $N(0, \sigma^2)$ distribution. The model of order $p_1 = p_2 = 1$ is described in Basu and Reinsel (1993). In the present paper we restrict our interest on this spatial structure, i.e. only of order 1. The assumed structure can be seen in Figure 1. One can see that for each point, only three neighbouring points coming from the upper left direction affect the point, implying a restricted spatial structure. In other words the spatial effect is propagated from left to the right and from top to the bottom only.

2.1 The stationary model

Ghodsi et al. (2012) extended the model for count data. Hence, they defined that the spatial observation located at site coordinate (i, j) follows an auto-regressive equation of the form:

$$Y_{ij} = \alpha_1 \circ Y_{i-1,j} + \alpha_2 \circ Y_{i,j-1} + \alpha_3 \circ Y_{i-1,j-1} + \epsilon_{ij}, \quad (1)$$

where $i = 1, \dots, n_1$, $j = 1, \dots, n_2$.

The dependence of Y_{ij} on its neighbours as defined by set S_{ij} is handled in equation (1) through the binomial thinning operator “ \circ ”. The binomial thinning mechanism emanates from the work of Steutel and Harn (1979) (see also Scotto, Weiß and Gouveia, 2015) for a summary of such operators) and is expressed as:

$$\alpha \circ Y = \sum_{s=1}^Y B_s(\alpha), \quad (2)$$

where $\alpha \in [0, 1]$, and $B_s(\alpha)$, $s = 1, \dots, Y$ are identically and independently distributed Bernoulli r.v with $P(B_s(\alpha) = 1) = 1 - P(B_s(\alpha) = 0) = \alpha$. In parsimony, we impose the assumption of independent thinning operator (Du and Li, 1991; Bu, McCabe and Hadri, 2006). In (1), $\{\epsilon_{ij}\}_{i=1, \dots, n_1, j=1, \dots, n_2}$ represents the corresponding innovation sequence of independent non-negative integer-valued random variables with finite mean λ_ϵ and finite variance τ_ϵ^2 and has a distributional form as $P_{\epsilon_{ij}}(\cdot)$. Furthermore, at any position (i, j) , ϵ_{ij} is assumed to be independent of all $Y_{i-k, j-\ell}$. In this simple form, the model in (1) is stationary if

$$\alpha_1 + \alpha_2 + \alpha_3 < 1. \quad (3)$$

Properties of this SINAR(1) model as well as estimation can be found in Ghodsi et al. (2012) and Ghodsi (2015). The process in equation (1) is proven to be ergodic in Ghodsi (2015) and Markovian in Pickard (1980).

2.2 The non-stationary model

Here, we extend the model to the non-stationary case by allowing site specific covariates to influence the mean of the innovation process. We denote the mean and variance of the innovation term as λ_{ij} and $\tau_{\epsilon_{ij}}^2$, respectively. We further consider λ_{ij} as a function of some position-variant and invariant covariates i.e. $\lambda_{ij} = f(x_{ij}^\top \beta)$ with $x_{ij} = [x_{ij1}, \dots, x_{ijp}]^\top$ and regression coefficients $\beta = [\beta_0, \dots, \beta_p]^\top$. Note that a log function is a standard choice for such cases leading to

$$\log \lambda_{ij} = x_{ij}^\top \beta.$$

We name this model as Generalized SINAR of order 1 (GSINAR(1)).

From the following binomial thinning properties,

$$\begin{aligned} E(\alpha \circ Y) &= \alpha E(Y) \\ V(\alpha \circ Y) &= \alpha(1 - \alpha)E(Y) + \alpha^2 V(Y) \\ \text{Cov}(\alpha_1 \circ Y_1, \alpha_2 \circ Y_2) &= \alpha_1 \alpha_2 \text{Cov}(Y_1, Y_2), \alpha_j \in (0, 1), j = 1, 2 \end{aligned}$$

and then we get the unconditional expectation of Y_{ij} to be

$$E(Y_{ij}) = \mu_{ij} = \alpha_1 \mu_{i-1,j} + \alpha_2 \mu_{i,j-1} + \alpha_3 \mu_{i-1,j-1} + \lambda_{ij}. \quad (4)$$

For the variance we get

$$\begin{aligned} V(Y_{ij}) = \sigma_{ij}^2 &= \alpha_1(1 - \alpha_1)\mu_{i-1,j} + \alpha_1^2 \sigma_{i-1,j}^2 \\ &+ \alpha_2(1 - \alpha_2)\mu_{i,j-1} + \alpha_2^2 \sigma_{i,j-1}^2 \\ &+ \alpha_3(1 - \alpha_3)\mu_{i-1,j-1} + \alpha_3^2 \sigma_{i-1,j-1}^2 \\ &+ 2\alpha_1 \alpha_2 \text{Cov}(Y_{i-1,j}, Y_{i,j-1}) \\ &+ 2\alpha_1 \alpha_3 \text{Cov}(Y_{i-1,j}, Y_{i-1,j-1}) \\ &+ 2\alpha_2 \alpha_3 \text{Cov}(Y_{i,j-1}, Y_{i-1,j-1}) + \tau_{\epsilon_{ij}}^2. \end{aligned} \quad (5)$$

By letting $\gamma(k, \ell) = \text{Cov}(Y_{i-k,j-\ell}, Y_{ij})$, we obtain a difference equation of the form:

$$\gamma(k, \ell) = \alpha_1 \gamma(k-1, \ell) + \alpha_2 \gamma(k, \ell-1) + \alpha_3 \gamma(k-1, \ell-1). \quad (6)$$

Closed form expressions for the above moments are difficult to obtain under non-stationary conditions. Whilst, unless assuming weak-stationarity, that is, ϵ_{ij} has constant mean λ_ϵ and variance τ_ϵ^2 , we obtain simple expression for the mean and variance with the covariances obtained by solving the difference equation $\gamma(k, \ell)$ using the approach in Basu and Reinsel (1993). However, the derivation of the covariance structure in equation (6) is not required in the estimation of the model parameters when using the conditional maximum likelihood equation illustrated as follows. Conditional on the neighbourhood S_{ij} , the probability mass function for the GSINAR(1) model is given by:

$$P(Y_{ij}|S_{ij}) = \sum_{s_1=0}^{R_1} \sum_{s_2=0}^{R_2} \sum_{s_3=0}^{R_3} p_{\alpha_1}(s_1|Y_{i-1,j}) p_{\alpha_2}(s_2|Y_{i,j-1}) p_{\alpha_3}(s_3|Y_{i-1,j-1}) P_{\epsilon_{ij}}(Y_{ij} - s_1 - s_2 - s_3), \quad (7)$$

where $R_1 = \min\{Y_{i-1,j}, Y_{ij}\}$, $R_2 = \min\{Y_{i,j-1}, Y_{ij} - s_1\}$ and $R_3 = \min\{Y_{i-1,j-1}, Y_{ij} - s_1 - s_2\}$ and $p_\alpha(s|Y) = \binom{Y}{s} \alpha^s (1-\alpha)^{Y-s}$; $s = 0, 1, \dots, Y$, i.e. the probability mass function of a binomial distribution. In the present paper we have considered different choices for $P_{\epsilon_{ij}}(\cdot)$. The standard assumption of a Poisson distribution limits the variability we expect in the data (see Appendix). A natural improvement is to consider mixed Poisson alternatives. We consider the negative binomial, Poisson-Inverse Gaussian as well as the Poisson-Lindley in order to allow for quite different effects. Also, we consider the COM-Poisson distribution in order to allow for a general model which accounts for underdispersion if we need so. We postpone the details until section 3.

Therefore, the log conditional maximum likelihood (CML) equation is then given by:

$$\ell(\theta) = \log L(\theta) = \sum_{i=2}^{n_1} \sum_{j=2}^{n_2} \log P(Y_{ij}|S_{ij}), \quad (8)$$

where $\theta = [\alpha_1, \alpha_2, \alpha_3, \beta, \nu]$, ν refers to the dispersion parameter of the innovation distribution if it exists and β is the vector of regression coefficients for the mean of the innovation. It can be seen in (Ghods, 2015) that

$$\hat{\theta} - \theta \sim N(0, I^{-1}(\theta)),$$

where $I(\theta)$ is the Hessian matrix. The CML equation in (8) is then maximized.

Some computational details are the following. We have used the `optim` function in R. Note that the conditional distribution needs to derive the convolution of three binomials plus the distribution of the innovation term. This can be computationally intensive. We have reduced the computational burden by observing that the probabilities of the binomial distribution are just the coefficients of a polynomial of order N where N is the number of trials in the binomial. As such computing the convolution of two binomial is equivalent to multiply two polynomials for which there are very fast procedures, like those in the library `pracma` in R. This reduced the computational effort and improved with respect to the errors produces by huge finite summations. Overall, maximization of (8) was rather simple even for complicated innovation distributions.

2.3 Related models

The general model in (1) has three parameters to introduce the spatial correlation, namely α_1 , α_2 and α_3 that described the vertical, horizontal and diagonal dependence respectively. One may eliminate some of the effects by setting the corresponding α parameter equal to 0. For example setting $\alpha_3 = 0$ we assume no-upper diagonal effect, while setting $\alpha_2 = 0$ we assume no horizontal effect. Such submodels can be very useful in order to examine and interpret the underlying situation for a dataset. For example, we may test and recognize which effect the vertical or horizontal is more important.

Another way to simplify the model is by assuming one common effect using the same parameter α for all directional relationships. Such a model takes the following form due to the properties of thinning operators:

$$Y_{ij} = \alpha \circ \sum_{(k,\ell) \in S_{ij}} Y_{k,\ell} + \epsilon_{ij}. \quad (9)$$

The model assumes that all neighbouring sites contribute the same to the structure. Such a model resembles simple INAR(1) time series models. It has the advantage of having less parameters to estimate and explain; all neighbours contribute the same. On the other hand, this may be restrictive since the spatial effects may differ due to direction and thus the model may fail to capture them correctly.

Model (9) allows for easy extensions to a general neighbouring structure. It is evident that by considering the set S_{ij} defining the neighbouring sites, this model can be generalized to a large extend including the non-regular lattice case which is more realistic in many applications. The model just assumes that all neighbours contribute to the observation at hand. Properties of such models as well as estimation is straightforward based on the results of the current section.

Finally, in the present paper we assume that the covariates enter in the model by the mean of the innovations. One may consider that spatial correlation parameters α_j may relate to some covariate information through a logit link function. For example, we can assume that $\text{logit}(\alpha_{1ij}) = x_{ij}^\top \delta_1$, where x_{ij} is some covariate information for the site (i, j) and δ_1 some vector of regression coefficients. In this case we assume that each point in space has a different spatial effect α_1 depending on some characteristics x_{ij} . For example, we may assume that altitude can change the spatial effect, which makes sense if we measure for example something which can be altered due to wind conditions. We believe that such a model, while it has some potential is special cases, it can complicate the model interpretation, especially if we have regression effects in both the mean and the autocorrelation parameters.

3 Simulation study

In this section, we perform simulation experiments, using equation (1), with different innovation distributions namely the Poisson, Negative Binomial (NB), Poisson-Lindley (PL), Conway-Maxwell Poisson (COM-Poisson / CMP) and Poisson-Inverse Gaussian (PIG) with rate or mean parameters commonly indicated by a link predictor $\lambda_{ij} = \exp(x_{ij}^T \beta)$ and dispersion index ν . As it is described in more details in the Appendix, using a distribution for the innovations that allows over/under dispersion, we also extend such properties to the observed spatial distribution and hence more realistic and flexible fitting can be achieved. In the present paper we consider the following distributions for the innovations.

- For Poisson innovations we assume that

$$P(\epsilon_{ij}) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{\epsilon_{ij}}}{\epsilon_{ij}!}; \epsilon_{ij} = 0, 1, \dots; \lambda_{ij} \geq 0.$$

- For NB innovations we use the following parametrization:

$$P(\epsilon_{ij}) = \frac{\Gamma(\nu + \epsilon_{ij})}{\Gamma(\nu) \epsilon_{ij}!} \left(\frac{\lambda_{ij}}{\lambda_{ij} + \nu} \right)^\nu \left(\frac{\nu}{\lambda_{ij} + \nu} \right)^{\epsilon_{ij}}; \epsilon_{ij} = 0, 1, \dots; \lambda_{ij} \geq 0, \nu \geq 0.$$

For this parametrization the mean is λ_{ij} and the variance $\lambda_{ij} + \lambda_{ij}^2/\nu$.

- For PL innovations we use

$$P(\epsilon_{ij}) = \frac{\lambda_{ij}^2 (\epsilon_{ij} + \lambda_{ij} + 2)}{(\lambda_{ij} + 1)^{\epsilon_{ij} + 3}}; \epsilon_{ij} = 0, 1, \dots; \lambda_{ij} > 0.$$

The mean is $(\lambda_{ij} + 2)/(\lambda_{ij}(\lambda_{ij} + 1))$ while the variance is

$$\frac{\lambda_{ij}^3 + 4\lambda_{ij}^2 + 6\lambda_{ij} + 2}{\lambda_{ij}^2(\lambda_{ij} + 1)^2}$$

PL can have different shapes than the other Poisson mixtures like the NB and PIG models.

- For COM-Poisson innovations we use

$$P(\epsilon_{ij}) = \frac{\lambda_{ij}^{\epsilon_{ij}}}{(\epsilon_{ij})!^\nu} \frac{1}{Z(\lambda_{ij}, \nu)}; \epsilon_{ij} = 0, 1, \dots; j = 0, 1, \dots; \lambda_{ij} \geq 0; \nu \geq 0,$$

where $Z(\lambda_{ij}, \nu) = \sum_{j=0}^{\infty} \frac{\lambda_{ij}^j}{j!^\nu}$. Note that λ_{ij} is not the mean of the distribution; the mean is hard to be written in closed form, but it is approximated by $\lambda_{ij}^{1/\nu}$.

- For PIG innovations we use the parameterization from the package *actuar* in R. Namely the pmf is defined with parameters λ and dispersion ν as

$$P(\epsilon_{ij}) = \left(\frac{2\nu}{\pi}\right)^{1/2} \exp\left(\frac{\nu}{\lambda_{ij}}\right) \left(\frac{a_{ij}}{\nu}\right)^{-(x-\frac{1}{2})} K_{x-\frac{1}{2}}(a_{ij}), \quad x = 0, 1, \dots, \quad \lambda_{ij}, \nu > 0,$$

where $a_{ij}^2 = 2\nu \left(1 + \frac{\nu}{2\lambda_{ij}^2}\right)$ and $K_x(a)$ is the modified Bessel function of the third kind. The mean is λ_{ij} and the variance is $\lambda_{ij} + \lambda_{ij}^3/\nu$.

The choice of the distributions for the innovation term attempts to cover a range of possible models. So, we have used the Poisson distribution as a starting point, two of the most famous mixed Poisson ones (negative binomial and Poisson-Inverse Gaussian), the COM-Poisson to allow for under-dispersion as well and finally the Poisson-Lindley since this is a tractable mixed Poisson distribution with very different shapes.

In order to simulate the grid we followed the following approach: We added an additional row and column with all values equal to 0, i.e. we set $Y_{0j} = Y_{i0} = 0$ for all i and j . Then we simulated the grid Y_{ij} , $i = 1, \dots, n_1 + 10$ and $j = 1, \dots, n_2 + 10$ based on the model in (1) using the chosen innovation distribution. Then we rejected the rows and columns from 0 up to 10 so as to keep the grid $n_1 \times n_2$.

3.1 Numerical Results: No covariates

For scenario 1, the simulation study assumes the following combinations of $(\alpha_1, \alpha_2, \alpha_3, \lambda, \nu)$ and grids:

1. C1: (0.35, 0.15, 0.2, 5, 0.5) and grid 25 x 25
2. C2: (0.25, 0.25, 0.3, 3, 0.8) and grid 40 x 40
3. C3: (0.6, 0.2, 0.15, 7, 2) and grid 50 x 50

Note that for Poisson-Lindley and COM-Poisson cases parameter λ is not the mean while ν has a different interpretation. So, in the simulations this is the value used to simulate the data. For Poisson, negative binomial and Poisson-Inverse Gaussian, λ is the mean and ν is the dispersion parameters, equal to 1 for the Poisson. Obviously for $\nu \rightarrow \infty$ we get the Poisson distribution in such cases.

For each scenario 1000 replications were obtained. The simulated mean estimates, their biases, root mean square errors (RMSEs) and standard deviations (SDs) are reported. The results are displayed in Tables 1, 2 and 3.

Table 1: Mean, Bias, RMSE and SD of estimates under different innovations under C1. Note that for Poisson-Lindley and COM-Poisson case parameter λ is not the mean. In the simulations this is the value used to simulate the data.

Innovation	Estimates	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\lambda}$	$\hat{\nu}$
Poisson	Mean	0.3503	0.1485	0.2018	4.9558	
	Bias	0.0003	-0.0015	0.0018	-0.0442	
	RMSE	0.0360	0.0425	0.0438	0.7580	
	SD	0.0350	0.0405	0.0338	0.7571	
NB	Mean	0.3555	0.1502	0.2178	5.5005	0.5660
	Bias	0.0055	0.0002	0.0178	0.5005	0.0660
	RMSE	0.0521	0.0601	0.0489	0.7280	0.4180
	SD	0.0355	0.0436	0.0426	0.6597	0.5296
PL	Mean	0.3477	0.1466	0.1969	5.0892	
	Bias	-0.0023	-0.0034	-0.0031	0.0892	
	RMSE	0.0367	0.0538	0.0715	0.7141	
	SD	0.0369	0.0541	0.0719	0.7233	
CMP	Mean	0.3557	0.1478	0.2035	4.2440	0.5770
	Bias	0.0057	-0.0022	0.0035	-0.7560	0.0770
	RMSE	0.0385	0.0502	0.0485	0.8524	0.5912
	SD	0.0381	0.0402	0.0414	0.6915	0.5996
PIG	Mean	0.3414	0.1416	0.2101	4.8164	0.4990
	Bias	-0.0086	-0.0084	0.0101	0.1836	-0.0010
	RMSE	0.0345	0.0350	0.0490	0.7839	0.5813
	SD	0.0399	0.0111	0.0261	0.6781	0.5793

Table 2: Mean, Bias, RMSE and SD of estimates under different innovations under C2. Note that for Poisson-Lindley and COM-Poisson case parameter λ is not the mean. In the simulations this is the value used to simulate the data.

Innovation	Estimates	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\lambda}$	$\hat{\nu}$
Poisson	Mean	0.2500	0.2489	0.2989	3.0290	
	Bias	0.0000	-0.0011	-0.0011	0.0290	
	RMSE	0.0010	0.0009	0.0011	0.0172	
	SD	0.0008	0.0007	0.0010	0.0152	
NB	Mean	0.2497	0.2505	0.3012	3.0389	0.7990
	Bias	-0.0003	0.0005	0.0012	0.0389	-0.0010
	RMSE	0.0218	0.0238	0.0294	0.6433	0.2543
	SD	0.0118	0.0218	0.0154	0.6354	0.2891
PL	Mean	0.2491	0.2479	0.2999	2.9990	
	Bias	-0.0009	-0.0021	-0.0001	-0.0010	
	RMSE	0.0216	0.0216	0.0245	0.2654	
	SD	0.0206	0.0211	0.0204	0.2655	
CMP	Mean	0.2487	0.2481	0.3009	3.0266	0.8009
	Bias	-0.0023	-0.0019	0.0009	0.0266	0.0009
	RMSE	0.0223	0.0227	0.0239	0.4625	0.4728
	SD	0.0222	0.0225	0.0209	0.4619	0.4731
PIG	Mean	0.2480	0.2488	0.2979	2.9901	0.8111
	Bias	-0.0020	-0.0012	-0.0021	-0.0099	0.0111
	RMSE	0.0183	0.0176	0.0210	0.3014	0.4467
	SD	0.0182	0.0106	0.0209	0.3004	0.4807

Table 3: Mean, Bias, RMSE and SD of estimates under different innovations under C3. Note that for Poisson-Lindley and COM-Poisson case parameter λ is not the mean. In the simulations this is the value used to simulate the data.

Innovation	Estimates	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\lambda}$	$\hat{\nu}$
Poisson	Mean	0.5997	0.2070	0.1495	6.9875	
	Bias	-0.0003	0.0070	-0.0005	-0.0125	
	RMSE	0.0009	0.0003	0.0010	0.0101	
	SD	0.0005	0.0003	0.0009	0.0100	
NB	Mean	0.5996	0.2004	0.1511	6.9973	2.0010
	Bias	-0.0004	0.0004	0.0011	-0.0027	0.0010
	RMSE	0.0211	0.0120	0.0151	0.0910	0.2170
	SD	0.0218	0.0140	0.0170	0.0987	0.2281
PL	Mean	0.6067	0.2016	0.1503	6.9898	
	Bias	0.0067	0.0016	0.0003	-0.0102	
	RMSE	0.0207	0.0140	0.0133	0.0119	
	SD	0.0195	0.0134	0.0123	0.0102	
CMP	Mean	0.5982	0.1918	0.1529	7.0131	2.0210
	Bias	-0.0018	-0.0082	0.0029	0.0131	0.0210
	RMSE	0.0210	0.0130	0.0214	0.3798	0.3720
	SD	0.0214	0.0134	0.0215	0.3898	0.3731
PIG	Mean	0.5991	0.1992	0.1499	6.9997	1.9830
	Bias	-0.0009	-0.0008	-0.0001	-0.0003	-0.0170
	RMSE	0.0151	0.0116	0.0150	0.0998	0.5430
	SD	0.0152	0.0115	0.0154	0.0950	0.4450

The simulation results illustrate that the estimates of the different parameters are consistent. This remark is noticed for all the SINAR with the different innovation distributions and under the different combinations of C1, C2, C3. The simulations also ensured that the estimates of the $\hat{\alpha}$'s satisfy the stability condition for stationarity given in (3).

Note that in all replications almost no problems to maximize the log-likelihood were detected. Some problems occurred in the COM-Poisson innovations. Problems are related to the built in functions dcomp and dcompoisson as they could not compute efficiently the normalizing constant $Z(\lambda, \nu)$ in the COM-Poisson implementations in few simulations.

3.2 Numerical Results: With covariates

For the case with covariates we have added a covariate, say, X for the different scenarios. So we assume for the innovations that

$$\log \lambda_{ij} = \beta_0 + \beta_1 X_{ij},$$

where the covariate X_{ij} was generated from a standard normal distribution. Again we have checked different grids, namely 30×30 , 50×50 and 80×80 to see how the size

of the grid scales up the variance and the biases (if any). We have used two scenarios:

- S1: $\alpha_1 = 0.15, \alpha_2 = 0.1, \alpha_3 = 0.2, \beta_0 = 0.6, \beta_1 = 0.5$,
- S2: $\alpha_1 = 0.05, \alpha_2 = 0.1, \alpha_3 = 0.05, \beta_0 = 0.1, \beta_1 = -0.5$.

One can see that the second scenario S2 has smaller spatial correlation parameters closer to the lower boundary and hence we would like to see the behaviour. Now we need to estimate all 5 parameters.

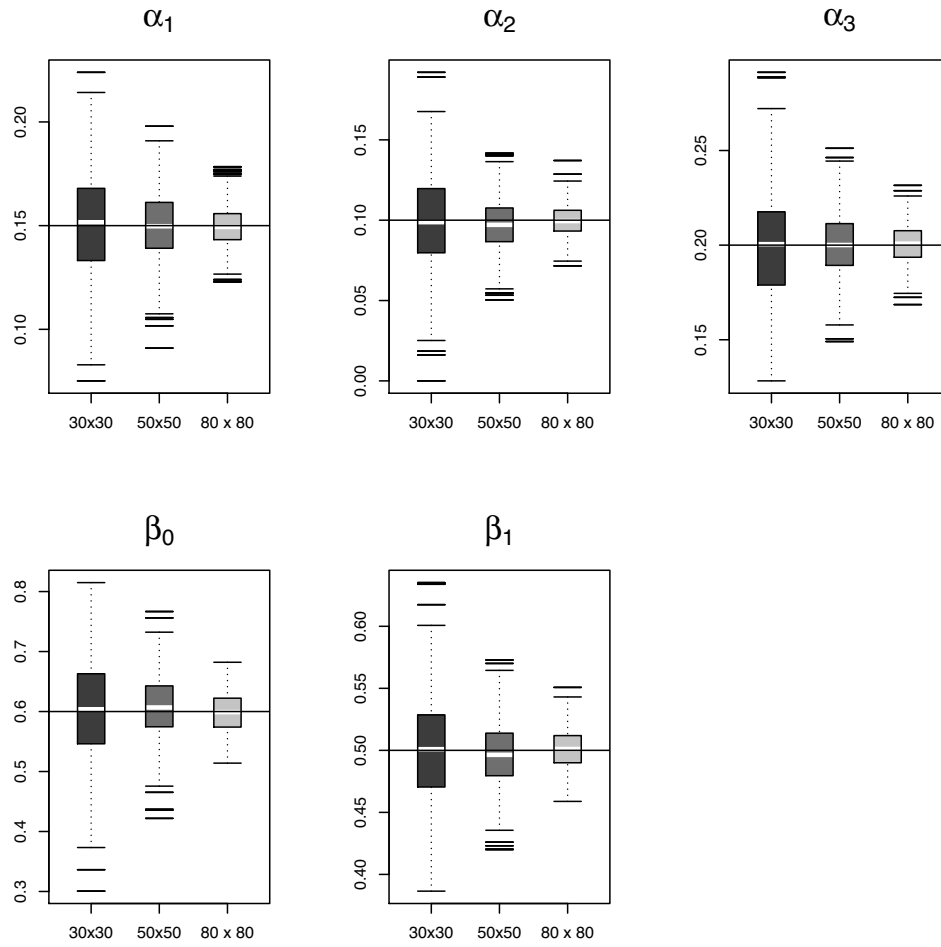


Figure 2: Boxplots for all five parameters under scenario S1 for the three different grids. The horizontal line represents the true value.

Figures 2 and 3 show the boxplots from 1000 replications under the two scenarios for the different grids. We present results from the Poisson innovations case only and similar

findings were obtained from the other models as well. The horizontal line represents the true value. One can see that even in the smaller grid the CML estimates correctly the true value. The variability as indicated by the boxplots reduces with the grid size as expected. Also from the boxplots one can see that the shape is symmetrical and confront with the asymptotic normality of the estimates.

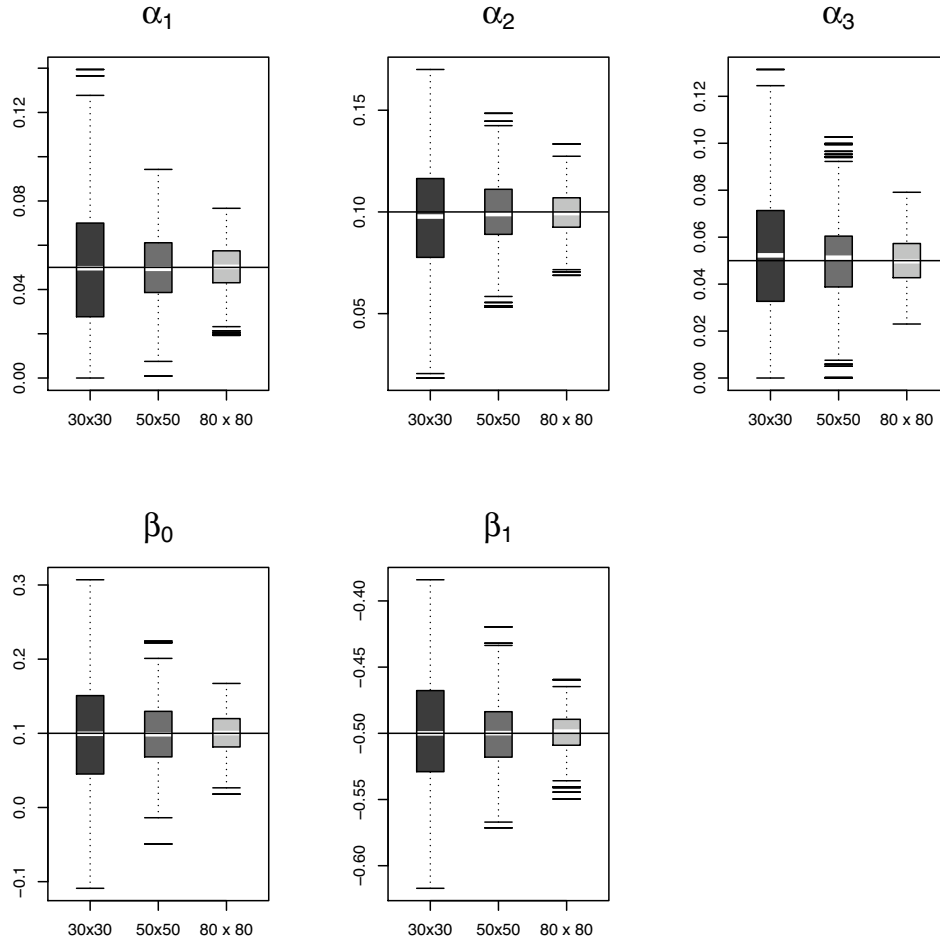


Figure 3: Boxplots for all five parameters under scenario S2 for the three different grids. The horizontal line represents the true value.

As far as the computational issues are concerned, no problems were found when fitting the models. As initial values we have used random value around the true ones, i.e. we simulated initial values by adding a uniform random variable in the interval $(-0.05, 0.05)$ to the true underlying values. For all runs we got convergence from optim function. To obey the restrictions of the parameter space we used transformations on the parameters.

4 Applications - Beilschmiedia data

4.1 The data

In studies of biodiversity of tropical rainforests, it is of interest to study whether the spatial patterns of the many different tree species can be related to spatial variations in environmental variables concerning topography and soil properties. Beilschmiedia dataset (Bei) in the `spatstat` package in R (Baddeley, Rubak and Turner, 2015) captures the locations of 3605 trees in a tropical rain forest. The data cover a $1000 \text{ m} \times 500 \text{ m}$ rectangular sampling region in the tropical rainforest of Barro Colorado Island. This data set is a part of a much larger data set containing positions of hundreds of thousands of trees belonging to hundreds of species. More details about the data can be found in Møller and Waagepetersen (2007).

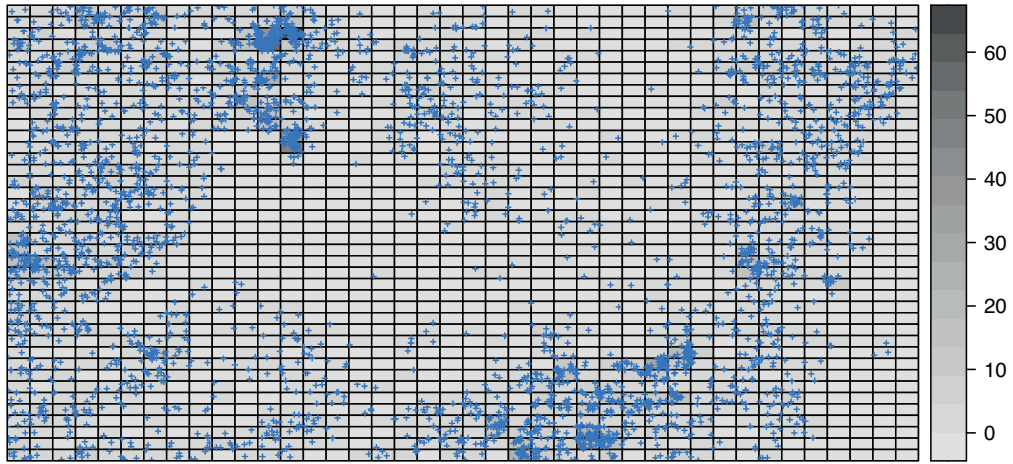


Figure 4: Spatial plot of Bei data.

A regular lattice of size 40×40 is created from the original dataset considering the number of trees inside each cell. The mean and variance are 2.25 and 12.4 respectively. The index of dispersion is 5.51 implying overdispersion that we cannot capture by Poisson innovations. Fitting models that allow for overdispersion is important. Figure 4 represents a spatial plot of the Bei data. In particular one can see the position of the trees in the lattice. The grey background depicts the observed counts and darker grey areas are those with more trees. In addition, Table 4 shows some values of the sample spatial autocorrelation of order k and ℓ for the Bei data. Here k refers to the horizontal direction and ℓ to the vertical direction. One can see that the horizontal autocorrelations are larger, supporting the use of a model like the one derived in section 2. The observed counts in the 40×40 grid can be also seen in Figure 5.

Table 4: Some values of the sample spatial autocorrelation for the Bei data.

ℓ	k				
	0	1	2	3	4
0	1.000	0.497	0.238	0.283	0.269
1	0.379	0.265	0.223	0.228	0.170
2	0.172	0.125	0.104	0.133	0.132
3	0.094	0.090	0.083	0.054	0.049
4	0.083	0.071	0.046	0.039	0.023

4.2 Results

To start with, we have fitted a series of models with different innovation distributions to capture the different aspects of the data. The fitted model to the Bei data using the CML estimation approach can be seen in Table 5.

Table 5: Comparison between different innovation distributions using Akaike information criterion (AIC): Application to BEI count data.

Innovations		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\lambda}$	$\hat{\nu}$	AIC
Poisson	Estimates	0.198	0.306	0.108	0.818		
	s.e	(0.019)	(0.013)	(0.011)	(0.034)		7179
NB	Estimates	0.151	0.225	0.053	1.209	0.205	
	s.e	(0.056)	(0.017)	(0.015)	(0.082)	(0.018)	5342
PL	Estimates	0.119	0.223	0.008	1.070		
	s.e	(0.016)	(0.017)	(0.015)	(0.039)		5882
CMP	Estimates	0.113	0.212	0.002	0.594	0.001	
	s.e	(0.017)	(0.018)	(0.015)	(0.028)	(0.042)	5732
PIG	Estimates	0.151	0.223	0.060	1.214	0.161	
	s.e	(0.015)	(0.016)	(0.015)	(0.097)	(0.022)	5343

It is observed from Table 5 that the SINAR(1) model using NB innovation distribution yields the lowest AIC value and hence outperforms the other models with different innovations. However, an interesting observation is that the PIG model has an AIC value which is really close to the selected model. In fact this implies that we need an overdispersed innovation distribution to capture the observed overdispersion. Note also that for all models the dependence parameters α_j are significant, supporting the usage of spatial models. We can also observe that the α_j are all positive values showing that geographically nearby values of the variable of interest are more similar than those of remote locations. Parameter α_2 that measures the horizontal dependence is larger. Perhaps this may relate to parameters associated with the lattice like the orientation with respect prevailing winds that expand the vegetation to some particular direction. For the COM-Poisson distribution, the model tends to a geometric distribution since parameter ν is almost zero. This may explain why the fit is not that good.

Table 6: Different models with negative binomial innovation distribution fitted to the Bei data. Models assumes different (or not ta all) spatial dependence. The full model with all three kind of unilateral effects is the chosen one.

Model	Param	Log-lik	AIC
No restriction	5	-2666.22	5337.442
M1: $\alpha_1 = \alpha_2 = \alpha_3$	3	-2686.20	5375.406
M2: $\alpha_3 = 0$	4	-2674.75	5353.490
M3: $\alpha_2 = 0$	4	-2765.51	5535.024
M4: $\alpha_1 = 0$	4	-2717.78	5439.554
M5: $\alpha_2 = \alpha_3 = 0$: only vertical	3	-2802.29	5607.586
M6: $\alpha_1 = \alpha_3 = 0$: only horizontal	3	-2746.06	5495.128
M7: $\alpha_1 = \alpha_2 = 0$: only diagonal	3	-2840.98	5684.954
no spatial effect $\alpha_1 = \alpha_2 = \alpha_3 = 0$	2	-2938.80	5879.606

Table 6 presents for the chosen negative binomial case some more spatial scenarios as mentioned in section 2.3. For example, model M1 assumes that all three α 's are the same, while models M2 to M4 that one of the spatial correlations is not present i.e. we set $\alpha_j = 0$ for $j = 1, 2, 3$ respectively. Actually we remove each time the vertical (M4), horizontal (M3) and the diagonal effects (M2). Finally, models M5 to M7 suggest that only one spatial effect suffices. The full model with all three kind of unilateral effects is the chosen one as judged by AIC, revealing the underlying structure of the data. One can see that the horizontal effect is larger as judged by the change in the LRT when we remove each effect.

In addition we use some covariate information available. The Bei data set is accompanied by covariate data giving the elevation (altitude) and slope of elevation in the study region. An important question arises is whether the intensity of Bei trees may be viewed as a spatially varying function of the covariates. We have fitted different models to examine the improvement offered by the covariates. Both covariates were found significant. The selected model can be seen in Table 7.

Table 7: Comparison between Poisson innovation and mixed Poisson innovations using Akaike information criterion (AIC): Application to Bei count data with both covariates.

Innovations		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\nu}$	AIC
Poisson	Estimates	0.187	0.286	0.102	-6.265	0.037	8.128		
	s.e	(0.012)	(0.013)	(0.011)	(0.671)	(0.004)	(0.478)		6940
NB	Estimates	0.146	0.219	0.055	-6.897	0.044	8.515	0.242	
	s.e	(0.016)	(0.017)	(0.016)	(1.507)	(0.010)	(1.327)	(0.089)	5293
PL	Estimates	0.115	0.208	0.000	4.904	-0.029	-6.839		
	s.e	(0.016)	(0.017)	(0.014)	(0.645)	(0.004)	(0.546)		5716
CMP	Estimates	0.111	0.202	0.006	-2.162	0.010	2.089	0.001	
	s.e	(0.016)	(0.018)	(0.015)	(0.116)	(0.001)	(0.236)	(0.031)	5626
PIG	Estimates	0.146	0.221	0.056	-15.053	0.097	16.539	0.210	
	s.e	(0.016)	(0.017)	(0.015)	(2.395)	(0.016)	(3.451)	(0.0213)	5294

It is observed from Table 7 that the SINAR(1) model using NB innovation distribution with covariates still yields the lowest AIC value and hence outperforms the other models with different innovations by producing much better fit to the data. PIG is very competitive to the NB.

We focus on the selected model with negative binomial innovations. From the results of Table 7 one can see that both covariates are statistically significant, with positive sign, hence increasing the altitude and the slope of elevation we obtain an increased number of trees, having adjusted for the effect of neighbouring areas. From the α 's we see that the larger effect comes from α_2 that measures the horizontal dependence. All spatial effects are statistically significant at 5%. The model implies a clear spatial dependence.

4.3 Goodness of Fit

In order to judge whether the fitted model is satisfactory we have worked a few ideas. To start with, we derived the one step ahead predictions based on the models. Namely, we derived for each data point

$$E(Y_{ij}|S_{ij}) = \hat{\alpha}_1 Y_{i-1,j} + \hat{\alpha}_2 Y_{i,j-1} + \hat{\alpha}_3 Y_{i-1,j-1} + \hat{\lambda}_{ij}$$

and

$$\log \hat{\lambda}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 X_{1ij} + \hat{\beta}_2 X_{2ij}$$

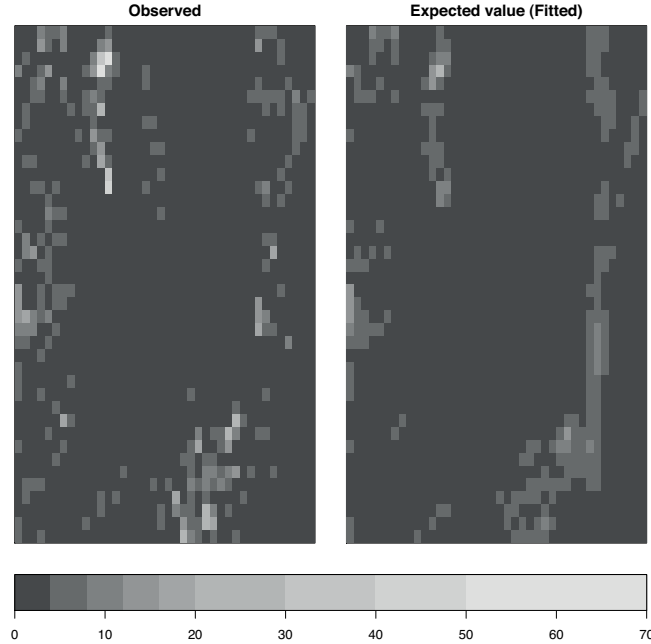


Figure 5: The observed counts in the 40×40 grid and the fitted based on the model. Fitted values are expected values based on the conditional expectation.

using the estimated parameters from the selected negative binomial model. The values can be seen in Figure 5 together with the observed counts. We emphasize that the predictions are the expected means that is why they cannot capture the extreme values. From the plots one can see that the model captures in a great extend the pattern.

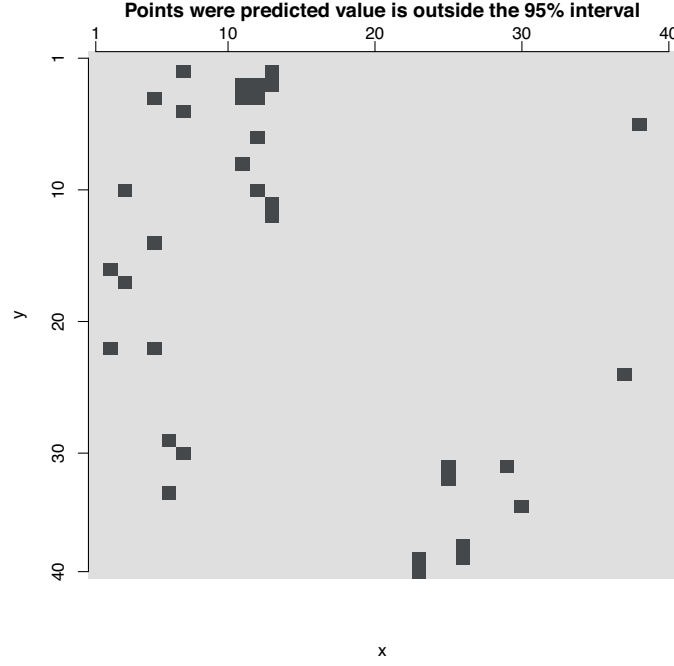


Figure 6: Points in the grid that the true value was outside the 95% prediction interval created.

To further exploit the quality of the predictions we have created for each data point (i, j) a 95% confidence interval for the prediction. To do so, we simulated 1000 values from the predictive probability mass function as provided in (7) and then based on them we created the intervals. Out of the 1521 values we predict (we did not predict the first row and column) only 33 (2.1%) values were outside of the interval, implying that the model was quite satisfactory. The values that lay outside the interval are depicted in Figure 6. One can see again that we have failed to predict some of the extreme values as one can see compared to the Figure 5.

Another important aspect of the model fitting lies on the ability of the model to capture the spatial dependence structure. To check this aspect we simulated grids of the same size 40×40 from the selected negative binomial using the estimated parameters. For each simulation we have estimated the spatial covariance at lags k and ℓ by

$$\hat{\gamma}(k, \ell) = \frac{1}{n} \sum_{i=k+1}^{40} \sum_{j=\ell+1}^{40} (Y_{ij} - \bar{Y})(Y_{i+k, j+\ell} - \bar{Y})$$

and then we derived the spatial correlation at lag k and ℓ as

$$\hat{\rho}(k, \ell) = \frac{\hat{\gamma}(k, \ell)}{s_Y^2},$$

where \bar{Y} and s_Y^2 are the mean and the variance estimated from the data.

We have used lags $k = 0, 1, 2$ and $\ell = 0, 1, 2$ and then we compared them with those values observed from the data in order to see whether the observed dependence structure could have been created by the model at hand. We show in Figure 7 the 95% confidence intervals created by 1000 simulations and the dot indicates the observed value. We see a good agreement. The two first values need perhaps improvement with a richer model but overall the model captures the underlying structure in a reasonable way.

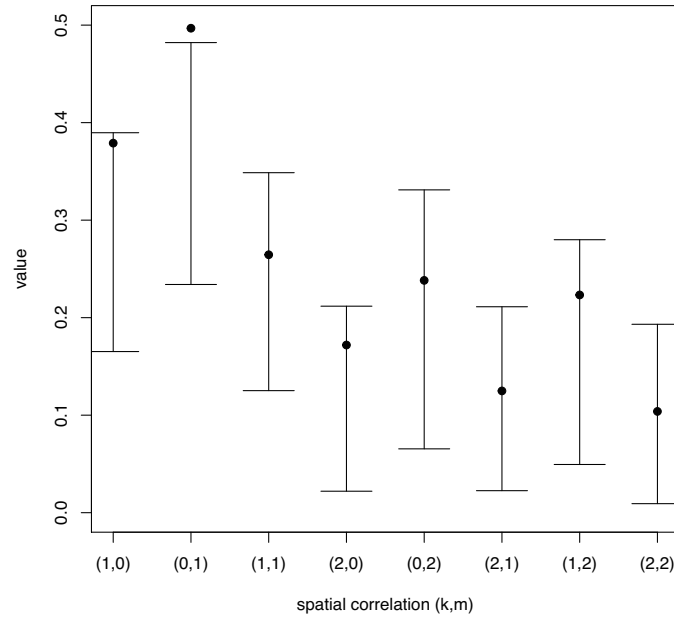


Figure 7: Points in the grid that the true value was outside the 95% prediction interval created.

As a summary, we believe that the current model fits satisfactorily the observed spatial structure.

5 Conclusion

This paper revisits and extends the simple stationary SINAR(1) model introduced by Ghodsi et al. (2012) and Ghodsi (2015). The SINAR(1) model is the first research in the modelling of the two dimensional unilateral spatial discrete data based on the thinning mechanism that allows to model explicitly the discrete nature of the data.

In the present paper we proposed some novel extensions of the existing SINAR(1) model. These novelties in fact overcome two important limitations of the simple SINAR(1). Firstly, in our model specification, we propose to model the data using overdispersed innovation distributions, while simultaneously allow covariate information to be used leading to a non-stationary model. While not treated in this version, one may also use offset in the regression part, like population size values, if needed. We also discuss parsimonious representations of the model at hand. The model parameters are estimated using the CML approach.

We acknowledge some restrictions of the current model which we consider to improve. The proposed model is based on the unilateral regular lattice case. One can extend the model to capture several other cases. For the regular lattice case, define the set of indices (k, ℓ) of the neighbouring observations for the (i, j) observation as S_{ij} . Then, in the general setting, the model can be written as for a general neighbourhood:

$$Y_{ij} = \sum_{(k, \ell) \in S_{ij}} \alpha_{k\ell} \circ Y_{k, \ell} + \epsilon_{ij},$$

where are usual, the ϵ_{ij} are the innovations. Defining appropriately the sets S_{ij} one can derive other models at the expense of parsimony.

Finally consider the typical case in spatial data where observations are indexed simply as Y_j to indicate the value at site j from a map with $j = 1, \dots, n$ sites, as for example the different regions of a country. Define as S_j the indices of its neighbours. In such case the model of order 1 can have the form:

$$Y_j = \alpha \circ \sum_{k \in S_j} Y_k + \epsilon_j$$

or equivalently if we define the $n \times n$ adjacency matrix W with elements w_{ij} with values equal to 1 if the sites i and j are neighbours and 0 otherwise, then we can write the models as:

$$Y_j = \alpha \circ \sum_{k \neq j} w_{kj} Y_k + \epsilon_j$$

to mimic typical order 1 models for spatial continuous data. Such generalization will be reported elsewhere.

Also note that in this paper we used only spatial model of first order. One may consider SINAR(p) models with higher order effects. Such extension needs special care. It is already known that simple INAR(p) models can have different interpretations /representations, (see the different approaches in Alzaid and Al-Osh (1990) and Jin-Guan and Yuan (1991)). Extending to a SINAR(p) model can have a large number of parameters making inference quite complex. Perhaps more parsimonious models like the one in Section 2.3 are easier to extend to higher order.

A Appendix section

Based on Ghodsi et al. (2012), for the case of a stationary model, we have for the marginal stationary mean μ_Y and the stationary variance σ_Y^2 that

$$\mu_Y = \frac{\mu_\epsilon}{1 - \alpha_1 - \alpha_2 - \alpha_3}$$

and

$$\sigma_Y^2 = \frac{\mu_Y \sum_{i=1}^3 \alpha_i (1 - \alpha_i) + \tau_\epsilon^2}{1 - (\alpha_1 + \alpha_2 \alpha_3) \lambda - (\alpha_2 + \alpha_1 \alpha_3) \eta - \alpha_3^2}$$

where $\eta = \frac{\alpha_2 + \alpha_3 \lambda}{1 - \alpha_1 \lambda}$ and

$$\lambda = \frac{(1 + \alpha_1^2 - \alpha_2^2 - \alpha_3^2) - \sqrt{(1 + \alpha_1^2 - \alpha_2^2 - \alpha_3^2)^2 - 4(\alpha_1 + \alpha_2 \alpha_3)^2}}{2(\alpha_1 + \alpha_2 \alpha_3)}$$

In the formulas, μ_ϵ and τ_ϵ^2 are the mean and the variance of the innovations respectively. Note a misprint in Ghodsi et al. (2012) for the variance. Define the index of dispersion $ID_Y = \sigma_Y^2 / \mu_Y$. Dividing the variance with the mean, we get for the index of dispersion of the spatial data that

$$ID_Y = \frac{\sigma_Y^2}{\mu_Y} = \frac{\sum_{i=1}^3 \alpha_i (1 - \alpha_i) + ID_\epsilon (1 - \alpha_1 - \alpha_2 - \alpha_3)}{1 - (\alpha_1 + \alpha_2 \alpha_3) \lambda - (\alpha_2 + \alpha_1 \alpha_3) \eta - \alpha_3^2}$$

which relates directly the index of dispersion of the innovation distribution ID_ϵ to that of the marginal, i.e. ID_Y . Since the denominator is positive and all the quantities in the nominator also are positive, an increase of ID_ϵ will lead to increase of the ID_Y . Thus assuming an overdispersed distribution for the innovations we can have much larger overdispersion in the observed spatial data.

One can see that even for the Poisson innovations the index of dispersion is larger than 1, however for reasonable values for counts this overdispersion is limited. The introduction of overdispersed innovations increase a lot the overdispersion as one can see in Figure 8. In Figure 8 the two axes depict the marginal mean and variance for a stationary model given above. The different lines correspond to different levels of overdispersion for the innovation distribution. We have used $\alpha_1 = \alpha_2 = \alpha_3 = 0.2$. The diagonal line refers to the case of equidispersion. Therefore, above that line we get overdispersion and below underdispersion. The red line ($ID=1$) corresponds to Poisson innovations. One can see that in this case we get small overdispersion for the spatial case. Increasing the overdispersion on the innovation, as for example considering a mixture of Poisson we get larger overdispersion. Note that an underdispersed innovation distribution, like the cases of COM-Poisson distribution, can lead to underdispersion.

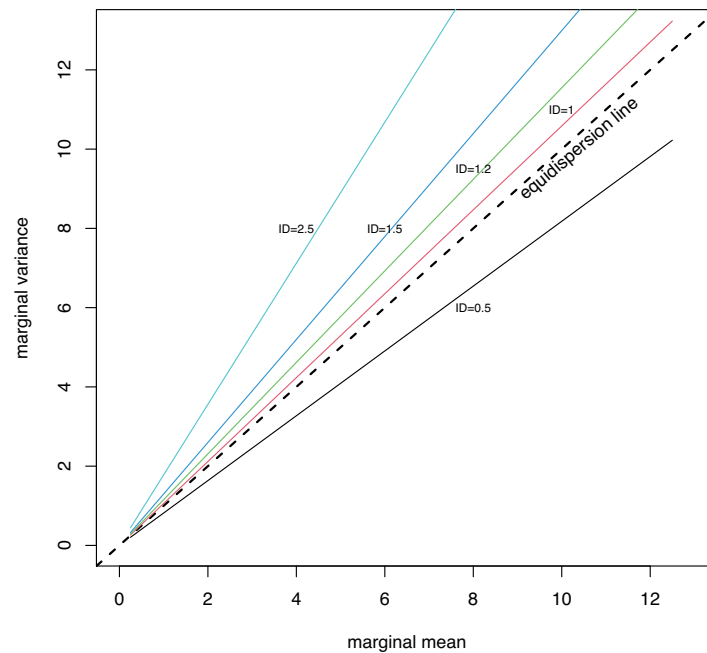


Figure 8: The marginal mean and variance for a stationary model. The different lines correspond to different levels of overdispersion for the innovation distribution. for the plot $\alpha_1 = \alpha_2 = \alpha_3 = 0.2$. The diagonal line refers to the case of equidispersion. ID implies the index of dispersion of the innovation distribution. One can see that for Poisson (red line) we get small overdispersion. Increasing the overdispersion of the innovations lead to increased overdispersion for the spatial distribution.

References

- Alvarez, M. (2020). Predicting traffic accident hotspots with spatial data science. <https://carto.com/blog/predicting-traffic-accident-hotspots-with-spatial-data-science/>
- Alzaid, A. and Al-Osh, M. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, 314-324.
- Baddeley, A., Rubak, E. and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman; Hall/CRC press.
- Basu, S. and Reinsel, G. C. (1993). Properties of the spatial unilateral first-order ARMA model. *Advances in Applied Probability*, 631-648.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36, 192-236.
- Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733-746.
- Besag, J., Mollié, A. and York, J. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1-20.
- Bu, R., McCabe, B. and Hadri, K. (2006). Maximum likelihood estimation of higher-order integer-valued autoregressive processes. *Journal of Time Series Analysis*, 29, 973-994.
- Chun, Y. (2014). An application to vehicle burglary. *Geographical Analysis*, 46, 165-184.
- Cressie, N. (1993). *Statistics for Spatial Data* (2nd ed.). New York: Wiley.
- Cressie, N. and Chan, N. (1989). Spatial modeling of regional variables. *Epidemiology and Infection*, 84, 393-401.

- Du, J. and Li, Y. (1991). The integer-valued autoregressive INAR(p) model. *Journal of Time Series Analysis*, 12, 129-142.
- Ghodsi, A. (2015). Conditional maximum likelihood estimation of the first-order spatial integer-valued autoregressive SINAR(1,1) model. *Journal of The Iranian Statistical Society*, 14, 15-36.
- Ghodsi, A., Shitan, M. and Bakouch, H. (2012). A first-order spatial integer-valued autoregressive SINAR(1,1) model. *Communications in Statistics- Theory and Methods*, 41, 2773-2787.
- Jin-Guan, D. and Yuan, L. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis*, 12, 129-142.
- Kruijer, W., Stein, A., Schaafsma, W. and Heijting, S. (2007). Analyzing spatial count data, with an application to weed counts. *Environmental and Ecological Statistics*, 14, 399-410.
- Lawson, A., Biggeri, A., Bohning, D., Lesaffre, E., Viel, J. and Bertollini, R. (1999). *Disease Mapping and Risk Assessment for Public Health*. Chichester, UK: Wiley.
- Lawson, A. and Williams, F. (2001). *An Introductory Guide to Disease Mapping*. New York: Wiley Medical Sciences.
- Mburu, L. and Bakillah, M. (2016). Modeling spatial interactions between areas to assess the burglary risk. *International Journal of Geo-Information*, 5, 47.
- McKenzie, E. (1986). Autoregressive moving-average processes with negative binomial and geometric marginal distributions. *Advanced in Applied Probability*, 18, 679-705.
- Møller, J. and Waagepetersen, R. (2007). Modern spatial point process modelling and inference (with discussion). *Scandinavian Journal of Statistics*, 34, 643-711.
- Obaromi, D. (2019). Spatial modelling of some conditional autoregressive priors in a disease mapping model: the Bayesian approach. *Biomedical Journal of Scientific and Technical Research*, 14.
- Pickard, D. (1980). Unilateral Markov fields. *Advances in Applied Probability*, 12, 655-671.
- Satria, R. and Castro, M. (2016). GIS tools for analyzing accidents and road design: A review. *Transportation Research Procedia*, 18, 242-247.
- Scotto, M., Weiß, C. and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: A review. *Statistical Modelling*, 15, 590-618.
- Steutel, F. and Harn, K. V. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7, 3893-899.
- Tevie, J., Bohara, A. and Valdez, R. (2014). Examination of the geographical variation in human west Nile virus: a spatial filtering approach. *Epidemiology and Infection*, 142, 2522-2529.
- Tjøstheim, D. (1983). Statistical spatial series modelling II: some further results on unilateral lattice processes. *Advances in Applied Probability*, 15, 562-584.
- Tjøstheim, D. (1978a). A measure of association for spatial variables. *Biometrika*, 65, 109-114.
- Tjøstheim, D. (1978b). Statistical spatial series modelling. *Advances in Applied Probability*, 10, 130-154.
- Tobler, W. R. (1969). Geographical filters and their inverses. *Geographical Analysis*, 1, 234-253.
- Valverde, J. and Jovanis, P. (2008). Analysis of road crash frequency with spatial models. *Transportation Research Record Journal of the Transportation Research Board*, 2061, 55-63.
- Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8, 158-183.

Information for authors and subscribers

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

SORT is an **Open Access** journal which **does not** charge publication **fees**.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. **SORT** strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)]]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- ☐ Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. *Journal of Computational and Graphical Statistics*, 7, 139–157.
- ☐ Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd Ed. Chapman & Hall / CRC, New York.
- ☐ Chapter in book: Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), *The Basel Risk Parameters: Estimation, Validation, and Stress Testing*. Springer, New York.
- ☐ Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. *Journal of Statistical Software*, 39 (13). <http://www.jstatsoft.org/v39/i13>. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75-100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions



The articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Spain License.

You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work), you may not use the work for commercial purposes and you may not alter, transform, or build upon the work.

Published articles represent the author's opinions; the journal SORT-Statistics and Operations Research Transactions does not necessarily agree with the opinions expressed in the published articles.

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76 – Fax +34-93.557.30.01
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form

SORT (Statistics and Operations Research Transactions)

Name _____
Organisation _____
Street Address _____
Zip/Postal code _____ City _____
State/Country _____ Tel. _____
Fax _____ NIF/VAT Registration Number _____
E-mail _____
Date _____
Signature _____

I wish to subscribe to ***SORT (Statistics and Operations Research Transactions)***) from now on

Annual subscription rates:

- Spain: €42 (4 % VAT included)
- Other countries: €46 (4 % VAT included)

Price for individual issues (current and back issues):

- Spain: €15/issue (4 % VAT included)
- Other countries: €17/issue (4 % VAT included)

Please send this subscription form (or a photocopy) to:

SORT (Statistics and Operations Research Transactions)
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-557 30 01

Or by e-mail to:

sort@idescat.cat