stats

MDPI

Article

A Family of Correlated Observations: From Independent to Strongly Interrelated Ones

Daniel A. Griffith

School of Economic, Political and Policy Sciences, University of Texas at Dallas, Richardson, TX 75080, USA; dagriffith@utdallas.edu; Tel.: +1-972-883-4950

Received: 22 May 2020; Accepted: 27 June 2020; Published: 30 June 2020

Abstract: This paper proposes a new classification of correlated data types based upon the relative number of direct connections among observations, producing a family of correlated observations embracing seven categories, one whose empirical counterpart currently is unknown, and ranging from independent (i.e., no links) to approaching near-complete linkage (i.e., n(n − 1)/2 links). Analysis of specimen datasets from publicly available data sources furnishes empirical illustrations for these various categories. Their descriptions also include their historical context and calculation of their effective sample sizes (i.e., an equivalent number of independent observations). Concluding comments contain some state-of-the-art future research topics.

Keywords: correlated data; social network series; space series; time series; space-time series

1. Introduction

Researchers in nearly every discipline regularly encounter correlated observations, which they also tend to label using adjectives like paired, matched, or dependent. Other vocabulary used to describe specialized versions of such data include clustered, repeated measures, longitudinal/panel (almost always, multiple repeated measures [1]), and cross-sectional. A principal feature of these data is that, unlike the uncorrelated observations situation, a covariance matrix for their observations is not diagonal (i.e., off-diagonal matrix entries are non-zero); awareness of this data aspect helps researchers avoid pseudo-replication complications [2]. One important property of data is that uncorrelated (i.e., a linear correlation of zero) observations imply their independence (i.e., their conditional and marginal distributions are the same) in general only when their joint distribution is Gaussian; not all uncorrelated random variables are independent. Meanwhile, Sainani [3] highlights an important property of correlated data: measurements for correlated observations tend to be more similar (positive associations) than measurements for uncorrelated observations (i.e., off-diagonal covariance matrix entries are zero). However, although rare in occurrence, this correlation property also generalizes to: measurements for correlated observations tend to be more dissimilar (negative associations) than measurements for uncorrelated observations. This dis/similarity data feature alludes to complicating impacts of correlated data on variances rather than arithmetic means. Not surprisingly, then, the simple concept of an arithmetic mean predates the more sophisticated notion of variance by more than a millennium, dating back to approximately 500BC, with Hunt naming the term in 1687 [4], Simpson showing that it minimizes the sum of squared deviations in 1755 [5], and Bowley showing that it is unbiased in 1897 [4,6]. Meanwhile, in one of his seminal papers, Fisher [7] p. 399, first used the statistical term variance, defining it as follows: the square of the standard deviation, because of the manner in which variances of independent random variables may be added [6]. This chronology indicates that correlated data escaped the attention of scholars for many centuries. Today it escapes the attention of most, if not all, introductory statistics textbook authors [8] (Chapter 7).

In deriving sample size formulae, Liu and Liang [9] enumerate a partial family of correlated data structures, apparently based upon types of numerical patterns in them: (1) uncorrelated (i.e., independent observations); (2) exchangeable (i.e., their order is unimportant to their joint probability distribution because all pairwise correlations of observations are the same—an equicorrelation covariance matrix; based upon the de Finetti probability theorem); (3) autoregressive (i.e., autocorrelation); and, (4) unstructured (i.e., as many as n(n-1)/2 different pairwise correlations exist). As its fundamental goal, this paper proposes and establishes a different classification scheme hinted at by statistical practice, especially in terms of residuals produced by regression techniques, one based upon sparsity of the inverse of a set of observations' covariance matrix: (1) uncorrelated; (2) matched k-tuples, $k \geq 2$; (3) time series autocorrelation; (4) spatial series autocorrelation; (5) space-time series autocorrelation; and, (6) network series autocorrelation.

Although recognized and conceptualized in the earlier days of the statistics discipline, difficulties attributable to correlated data were not topics of much study, and the timeline sequence for their recognition is not intuitive. The first formally acknowledged after uncorrelated observations was temporal autocorrelation by Laplace in the early 1800s, in terms of time series of daily barometric pressure [10], p. 151. Next was spatial autocorrelation in 1900 by Student, who commented that this category of correlated observations impacts upon the sampling distribution of the Pearson product moment correlation coefficient [11], p. 3. Then, thirty-one years later, Hotelling [12] recognized what seems like the more natural successor to independent observations, namely matched ones comprising n/2 pairs of uncorrelated observations (the presences of pairings results in n always being even; in general, $k \geq 2$ signifies that n is a multiple of k). Although Einstein was the first to show a space-time linkage, in 1909, supplanting Newton's notion of absolute space and time, Keller [13] appears to be the first to have penned the term space-time autocorrelation (according to Google Scholar and Web of Science searches); publications prior to his jointly treating space and time do so as though they are independent and separable. Today, the advent of contemporary network science [14] established recognition of much more highly interconnected observations, especially with its small world concept, a theme highly relevant to contemporary big data discussions.

The main point of this paper is to articulate a classification based upon the density of the inverse of observation covariance matrices (supporting the use of regression examples), whereas its purpose is to summarize, contrast with Liu and Liang's [9] categorization, and illustrate this newer classification for correlated data.

2. Salient Definitions and Notation

Let n denote the number of independent observations for uncorrelated data (i.e., the sample size), and n* denote the number of equivalent independent observations for correlated data (i.e., the effective sample size). The inverse covariance matrix for observations, denoted in general by n-by-n matrix $\mathbf{V}\sigma^{-2}$, where σ^2 is the common variance across a sample of n observations, may be written in terms of matrix $(\mathbf{I} - \rho\mathbf{C})$—where $\mathbf{I}$ is the n-by-n identity matrix and n-by-n matrix $\mathbf{C}$ is binary 0–1, with entries $c_{ij} = 1$ if distinct observations i and j are directly correlated, and $c_{ij} = 0$ otherwise; $c_{ii} = 0$—for a first-order (i.e., matrix $\mathbf{C}$ has an exponent of 1) or relatively small domain correlation structure, where ρ denotes the correlation parameter and $\mathbf{C}$ denotes the basic observations direct linkage structure, or $(\mathbf{I} - \rho\mathbf{C})^{\mathrm{T}}(\mathbf{I} - \rho\mathbf{C})$ for a second-order (i.e., matrix $\mathbf{C}$ has an exponent of 2) or larger domain correlation structure, where superscript T is the matrix transpose operator. This specification is the one employed for many, if not most, autoregressive model specifications. The density for a connected graph representation of matrix $\mathbf{C}$, which in this graph-theoretic setting usually is labelled an adjacency matrix, denotes the percentage of off-diagonal ones in it; this quantity is an index of the sparsity of matrix $\mathbf{C}$.

The multivariate normal probability model furnishes a foundation for understanding correlated data. It may be written, for a sample of size one with n observations (i.e., the n observations constitute

a vector of n random variables that are the same), and aligning with a linear regression model specification, as follows:

$$\frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{-1/2}\sigma^n} e^{-(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})/(2\sigma^2)} \tag{1}$$

where $\mathbf{Y}$ is a n-by-1 vector of response variables, $\mathbf{X}$ is a n-by-(p + 1) matrix of p covariates and a vector of ones, $\mathbf{1}$, for the intercept term, and $\boldsymbol{\beta}$ is a (p + 1)-by-1 vector of regression coefficients. If the n observations are uncorrelated (i.e., $\mathbf{V}^{-1} = \mathbf{I}$), then Equation (1) reduces to the likelihood of an independent and identically distributed (i.i.d.) sample of a single Gaussian variable:

$$\prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma} e^{-(y_i-\mathbf{x}_i\boldsymbol{\beta})^2/(2\sigma^2)} \tag{2}$$

the joint univariate probability density function for n independent observations. In the context of this independence case, the inverse covariance matrix has n(n − 1) off-diagonal entries of zero, and n* = n.

Focusing on the sampling distribution of the mean, normality based upon the Central Limit Theorem supports a formula for n*, where the linear regression equation is $\mathbf{Y} = \mu\mathbf{1} + \mathbf{V}^{-1/2}\boldsymbol{\varepsilon}$ for i.i.d. normal (N, which denotes the Gaussian probability distribution) errors $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2)$, σ_ε^2 is the error variance, and $\mathbf{0}$ denotes the null vector:

$$E\left(\hat{\sigma}\frac{y}{2}\right) = E\left[\frac{(\bar{y}-\mu)^2}{n}\right] = \frac{TR\left(\mathbf{V}^{-1}\right)}{TR\left(\mathbf{V}^{-1}\right)} E\left[\frac{\mathbf{1}^T\boldsymbol{\varepsilon}^T\mathbf{V}^{-1}\boldsymbol{\varepsilon}\mathbf{1}}{n^2}\right] = \frac{\frac{TR\left(\mathbf{V}^{-1}\right)}{n}\sigma_\varepsilon^2}{\frac{TR\left(\mathbf{V}^{-1}\right)}{\mathbf{1}^T\left(\mathbf{V}^{-1}\right)\mathbf{1}}n} \tag{3}$$

where E(•) denotes the calculus of expectations, $\hat{\sigma}\frac{y}{2}$ is the estimated variance of the sample mean, $\bar{y}$, and TR is the matrix trace operator. In Equation (3), $TR(\mathbf{V}^{-1})/n$, from the expected value of the sample variance $E(s^2) = [TR(\mathbf{V}^{-1})/n]\,\sigma_\varepsilon^2$, is the variance inflation factor associated with correlation amongst observations; if no correlation exists, this term reduces to 1 [i.e., $TR(\mathbf{I}) = n$]. The denominator of the right-hand side of Equation (3) defines the effective sample size:

$$n* = \frac{TR\left(\mathbf{V}^{-1}\right)}{\mathbf{1}^T\left(\mathbf{V}^{-1}\right)\mathbf{1}}n \tag{4}$$

A relevant standard mathematical statistics theorem [15] pertaining to variance (VAR) here is

Theorem 1. $VAR(Y_1 + Y_2) = VAR(Y_1) + VAR(Y_2) + 2\rho\sqrt{VAR(Y_1) \times VAR(Y_2)}$, *which, for* $VAR(Y_1) = VAR(Y_2) = \sigma^2$, *reduces to*

$$VAR(Y_1 + Y_2) = 2(1 + \rho)\,\sigma^2 \tag{5}$$

.

This theorem implies

$$VAR(Y_1 - Y_2) = 2(1 - \rho)\,\sigma^2 \tag{6}$$

An additional assumption of independent Gaussian observations (i.e., $\rho = 0$) reduces Equations (5) and (6) to $2\sigma^2$.

In summary, the main point of this section is the presentation of a common observational covariance matrix formulation consistent with standard regression model specifications.

3. A Family of Correlated Observations: Categories Illustrated with Empirical Examples

A sizeable mathematical statistical theory literature exists about independent observations. This section addresses the remaining divisions of the proposed correlated data classification scheme:

(2) matched k-tuples, k ≥ 2; (3) time series autocorrelation; (4) spatial series autocorrelation; (5) space-time series autocorrelation; and, (6) network series autocorrelation. One common pragmatic objective is a parsimonious parametrization of the correlation structure that optimizes the efficiency of parameter estimation [16], whose implementation most often is as a single covariation parameter characterizing the correlated observations constituting a sample; one exception to this preference is the frequent use of two parameters to describe space-time autocorrelation, which still is very parsimonious. This paper adopts this simplest of specifications for illustrative purposes. The arithmetic mean serves as a convenient statistical concept for some of the conceptual discussions and numerical illustrations presented in this section, which also exploit its historical and theoretical relationships with the variance.

3.1. Correlated Data: Matched K-Tuples, K ≥ 2

A preponderance of empirical evidence advocates that pairs (i.e., k = 2) are the most common, and most frequently occurring in history, matched k-tuple. Additionally, history discloses that the workhorse of classical statistics is linear regression, which offers a technique for calculating means and, for example, differences of two means. Consider the regression equation:

$$\mathbf{Y} = \mu\mathbf{1} + \beta(\mathbf{G}_1 - \mathbf{G}_2) + \mathbf{V}^{-1/2}\boldsymbol{\varepsilon} \tag{7}$$

where $\mathbf{G}_j$ is a n-by-1 indicator variable such that its cell entry is 1 if an observation is in group j (j = 1, 2), and 0 otherwise, $\mu_1 = \mu + \beta$, $\mu_2 = \mu - \beta$, the entries in vector $\mathbf{Y}$ are sequential pairs (i.e., y_i and y_{i+1} are matched, where i is an odd consecutive integer index), and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$. Here the observations covariance structure matrix $\mathbf{V}$ is given by

$$\begin{pmatrix} 1 & -\rho & 0 & & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{8}$$

which is a n-by-n block-diagonal matrix with n/2 2-by-2 matrices on its diagonal such that

$$\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \rho\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \Longrightarrow \mathbf{I}_{n/2} \otimes \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \rho\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right] \tag{9}$$

where ⊗ denotes the Kronecker product matrix operation, and $\mathbf{I}_{n/2}$ denotes a $\frac{n}{2}$-by-$\frac{n}{2}$ identity matrix. Because the quantity of interest involves the variance of the difference between (rather than the sum of) two random variables, matrix $\mathbf{V}$ is of the form

$$\begin{pmatrix} 1 & \rho & 0 & & 0 \\ \rho & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = \left[\mathbf{I}_{n/2} \otimes \begin{pmatrix} \sqrt{\frac{1+\rho}{2}} & \sqrt{\frac{1-\rho}{2}} \\ \sqrt{\frac{1+\rho}{2}} & -\sqrt{\frac{1-\rho}{2}} \end{pmatrix}\right] \times \left[\mathbf{I}_{n/2} \otimes \begin{pmatrix} \sqrt{\frac{1+\rho}{2}} & \sqrt{\frac{1+\rho}{2}} \\ \sqrt{\frac{1-\rho}{2}} & -\sqrt{\frac{1-\rho}{2}} \end{pmatrix}\right]. \tag{10}$$

Exploiting the block-diagonal form to write this latter matrix avoids Kronecker products:

$$\begin{pmatrix} 1 & \rho & 0 & & 0 \\ \rho & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = \mathbf{I} + \rho\begin{pmatrix} -\mathbf{I} + \mathbf{1}\mathbf{1}^{\mathrm{T}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & -\mathbf{I} + \mathbf{1}\mathbf{1}^{\mathrm{T}} \end{pmatrix}, \tag{11}$$

where the diagonal blocks are of dimension k-by-k. The generalized least squares estimate of β is $b = \frac{\bar{y}_1 - \bar{y}_2}{2}$, with a variance, s_b^2, of $\frac{1-\rho}{2\left(\frac{n}{2}\right)}\ \hat{\sigma}_\varepsilon^2$. The resulting t-test statistic is the customary $\frac{\bar{y}_1 - \bar{y}_2}{\hat{\sigma}_\varepsilon \sqrt{\frac{2(1-\rho)}{\frac{n}{2}}}}$, an outcome corroborated by the preceding mathematical statistics theorem for the difference between two random variables. The number of zeroes in both the inverse covariance matrix and its square root is $n(n-2)$; in other words, its network density is $100/(n-1)\%$; this quantity generalizes to $100(k-1)/(n-1)\%$, where $k \geq 2$ is the number of matched observations (e.g., repeated measures).

For this correlated observations case (the Liu-Liang exchangeability category), the effective sample size is

$$n^* = [(k-1)/(1+\rho) - (k-2)]n = n/(1+\rho) \text{ for } k = 2, \quad (12)$$

with the positive definite restriction of $-1/(k-1) < \rho < 1/(k-1)$; this interval is $(-1, 1)$ for $k = 2$, $(-1/2, 1/2)$ for $k = 3$, and continues to shrink toward $(0, 0)$ as k increases. In other words, a sufficiently large number of repeated measures, k, results in $n^* \approx n/k = kn_o/k = n_o$, the number of independent observations, often assumed in this situation by default. Regardless, if $\rho = 0$, then $n^* = n$, the case for the difference of two means for two independent samples with a common variance. If $\rho = 1$, then, for matched pairs, $n^* = n/2$, the case for the difference of two means for two matched/paired samples with a common variance. Because the ideal $k = 2$ study involves the same n/2 observations with a pair of repeated measures (e.g., before-after), presumably ρ never should be negative. Andrews and Herzberg ([17] https://www.york.ac.uk/depts/maths/histstat/pml1/r/andrews.htm); and the R Project furnish empirical examples of this type of data with their respective Table 13.1 and twins dataset (https://app.quadstat.net/dataset/r-dataset-package-kmsurv-twins). The R Project furnishes female-male twins data about age of death ($n = 24$, $r = 0.83$). Table 13.1 furnishes same day measures of wind direction made in the morning and in the evening at a given location ($n = 210$; $r = 0.38$, and density = 0.5%). The former t-test statistic increases from 0.74 (ignoring observational correlation; i.e., pseudo-replications [2]) to 1.67 (accounting for observational correlation), and the latter increases from 2.15 (ignoring observational correlation) to 2.74 (accounting for observational correlation). Figure 1 portrays these two cases. The respective n* values here are roughly 13 (rather than 24) and 152 (rather than 210). As an important aside, the variances in Figure 1a suggest the potential for nonconstant variance; however, Pitman's correlated variance t-test confirms that the two sample variances (i.e., 103.06 and 61.72; 66.11 pooled variance) are not significantly different ($t = 1.47$, $df = 10$).

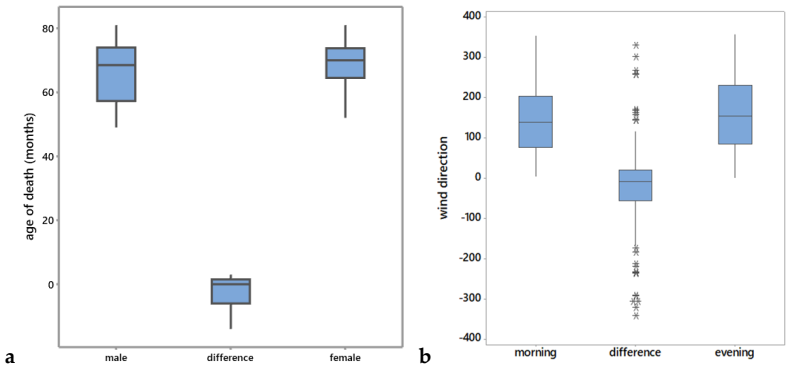

Figure 1. Boxplots for selected publicly available data. Left (**a**): R Project twins dataset results. Right (**b**): Results for Table 13.1 from Andrews and Herzberg [17].

This preceding conceptualization resembles that for analysis of variance (ANOVA) with repeated measures, which underscores how correlated observations impact upon variance calculations, with an assumption of a compound symmetric observations covariance matrix, which is consistent with positing a constant ρ. This implementation produces the following ANOVA tabulation for the twins dataset (assuming an equicorrelation covariance matrix):

Source	df	Mean Squared Error	Correlated data		Pseudo-Replication Data	
			F-Ratio	Probability	F-Ratio	Probability
Between	1	45.38	2.79	0.12	0.55	0.47
Within	22	82.39				
Error	11	16.28				
Twins	11	66.11				
Total	23					

The F-ratios here are the respective squared values of their preceding paired *t*-test statistics. This perspective asserts a sample size of n = 24, paralleling the preceding linear regression conceptualization. The network density in this case study is 4.3%.

One popular practice of observations matching is with monozygotic (i.e., identical) twins, especially those raised in the same shared environment rather than apart; in this situation, the correlation between, for example, IQ scores is approximately 0.86 [18]. Another familiar correlated data situation of this type is family members constituting a household because this assemblage of people engages in many activities as a group. A weaker matching practice is between treatment and control observations, where similarity of only a relatively small set of selected attributes sanctions the creation of artificial pairs. Andrews and Herzberg [17] include correlated data examples, such as artificial pairings (Tables 21.1, 27.2, 33.1, and 39.1), and same-observation repeated measures (Tables 24.1, 25.1, 28.3, 35.1, and 41.1). Hand et al. ([19] https://www2.stat.duke.edu/courses/Spring03/sta113/Data/Hand/Hand.html) include dozens of correlated data examples, ranging from the strengths of chemical pastes (Table 16), cork deposits by the compass direction sides of trees (Table 55), before-after patient treatment (Tables 72, 202, and 285), house insulation contrasts (Tables 88 and 93), linear and road distance separating locations (Table 115), chemical and magnetic measurements of iron in slag specimens (Table 132), and brothers' head sizes (Table 111) and heights of husbands and their wives (Table 231).

Repeated measures ANOVA extends the paired t-test to more than two means; in other words, k > 2. The utility of this correlated sample ANOVA technique is that it effectively removes extraneous variability attributable to pre-existing observational differences. The simplest block-diagonal version of the inter-observations correlation matrix for three means is as follows:

$$\begin{pmatrix} 1 & -\rho & -\rho & 0 & & 0 \\ -\rho & 1 & -\rho & 0 & \cdots & 0 \\ -\rho & -\rho & 1 & 0 & & 0 \\ 0 & 0 & 0 & 1 & & 0 \\ & \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = \mathbf{I}_{n/3} \otimes \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \rho \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \right], \quad (13)$$

where n always is a multiple of three, and which, as noted in the preceding discussion, is consistent with assuming a compound symmetric observations covariance matrix. For illustrative purposes, consider the distinct (i.e., triplet order is unimportant) primitive (i.e., the greatest common divisor is 1) Pythagorean quadruplets of Number Theory, each of which is a sum of triplets (i.e., a, b, and c) of squared integers that equals another squared integer (i.e., d): $a^2 + b^2 + c^2 = d^2$, $0 < a \leq b \leq c$, where the count of such numbers is 347 for $a \leq b \leq c \leq d \leq 100$ ([20] the dataset utilized here was extracted from https://pastebin.com/FGwtqsrs and then screened to delete non-distinct and non-primitive entries). This set of integers {a, b, c} constitutes matched triplets. Box-Cox transformations for normality coupled with division by appropriate quantities to adjust for unequal variances attributable to the ordering of and constraints on a, b, and c, yield the following analysis variables: tra = $\sqrt{a-1}$, trb = $[(b-2)^{0.75}]/3.2$, and trc = $[(c-2)^{1.25}]/43.1$. The resulting correlated sample variances are not statistically significantly different: $s^2_{tra} \approx 1.61^2$, $s^2_{trb} \approx 1.61^2$, and $s^2_{trc} \approx 1.60^2$. Furthermore, these three transformed integers'

pairwise correlations are $r_{tra,trb} = 0.52$, $r_{tra,trc} = 0.22$, and $r_{trb,trc} = 0.45$. The null hypothesis of interest here is:

$$H_0: \mu_{tra} = \mu_{trb} = \mu_{trc}. \quad (14)$$

The associated ANOVA tabulation is as follows, assuming equicorrelation with $\hat{\rho} = 0.40$ and a common variance (i.e., an assumption implied by the lack of three significantly different sample variances):

Source	df	Mean Squared Error	Correlated Data		Pseudo-Replication Data	
			F-Ratio	Probability	F-Ratio	Probability
Between	2	19.53	7.55	<0.001	12.51	<0.001
Within	1038	2.59				
Error	692	1.56				
Triplets	346	1.03				
Total	1040					

These findings imply that the means are different in the population, based upon the sample means of $\overline{tra} \approx 3.30$, $\overline{trb} \approx 3.76$, and $\overline{trc} \approx 3.43$; this outcome highlights the current concern about distinctions between statistical and substantive differences [21]. The network density for this case study is 0.2%, and $n^* = 450$; as noted previously, to ensure a positive-definite covariance matrix, the equicorrelation assumption restricts ρ in this repeated measures triplets case (i.e., $k = 3$) to the interval $(-1/2, 1/2)$.

In summary, the main point of this section is that repeated measures is a genre of correlated data whose regression residual specification is exactly the same in form as the other categories of correlated data. Another is that for sufficiently large k, it disappears from the Liu-Liang classification.

3.2. Correlated Data: Temporal Autocorrelation

Here matrix **V** includes, for a response variable vector **Y**, whose elements are organized in ascending time order, time structure matrix

$$C_T = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & & 0 & 0 \\ & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix} = \mathbf{I}_T - \rho \begin{pmatrix} \frac{1-\sqrt{1-\rho^2}}{\rho} & 0 & 0 & & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 \\ & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \quad (15)$$

where $n = T$, and $\mathbf{I}_T$ is a T-by-T identity matrix. Now the effective sample size is $n^* = n^2/[n + 2\sum_{k=1}^{n-1}(n-k)\rho^k]$, and the number of zeroes in the basic observations covariance structure matrix $\mathbf{C}_T$ is $(n-1)^2$. Again, if $\rho = 0$, then $n^* = n$, whereas if $\rho = 1$, then $n^* = 1$. Because a time series observation only interacts with its preceding observation(s), one-dimensional and one-directional temporal autocorrelation often is very strong and usually positive. The number of zeroes in the inverse covariance matrix is $(n-1)^2$; in other words, its network density is $100/n\%$.

Hand et al. [19] include Table 121 reporting annual whooping crane counts for 35 years (1938–1972; Figure 2a; $n = 35$, $\hat{\rho} = 0.95$, density = 2.9%). For this example, the sample variance estimate decreases from 12.3^2 (ignoring observational correlation) to 4.6^2 (accounting for observational correlation), and $n^* = 2$. Andrews and Herzberg [17] furnish Table 62.1, time series data tabulated for the quarterly size of the pig herd in the United Kingdom between 1967 and 1978 (Figure 2c; $n = 48$, $\hat{\rho} = 0.94$, density = 2.1%). For this example, again $n^* = 2$. Both of these publicly available collections of relatively small datasets furnish numerous additional examples of time series, some with weaker temporal autocorrelation (e.g., Table 8.1 in Andrews and Herzberg [17], days between coal mining disasters,

$\hat{\rho} = 0.31$), and some with the relatively rare case of negative temporal autocorrelation [e.g., Table 280 in Hand et al. [19], geyser eruption timings, $\hat{\rho} = -0.70$]. Because these specimen dataset sources are dated, many of their time series can be augmented with additional more recent observations (e.g., see https://www.canada.ca/en/environment-climate-change/services/species-risk-public-registry/cosewic-assessments-status-reports/whooping-crane-2010.html for an extension of the whooping crane time series to 2007; this data source contains a few discrepancies vis-à-vis Table 121).

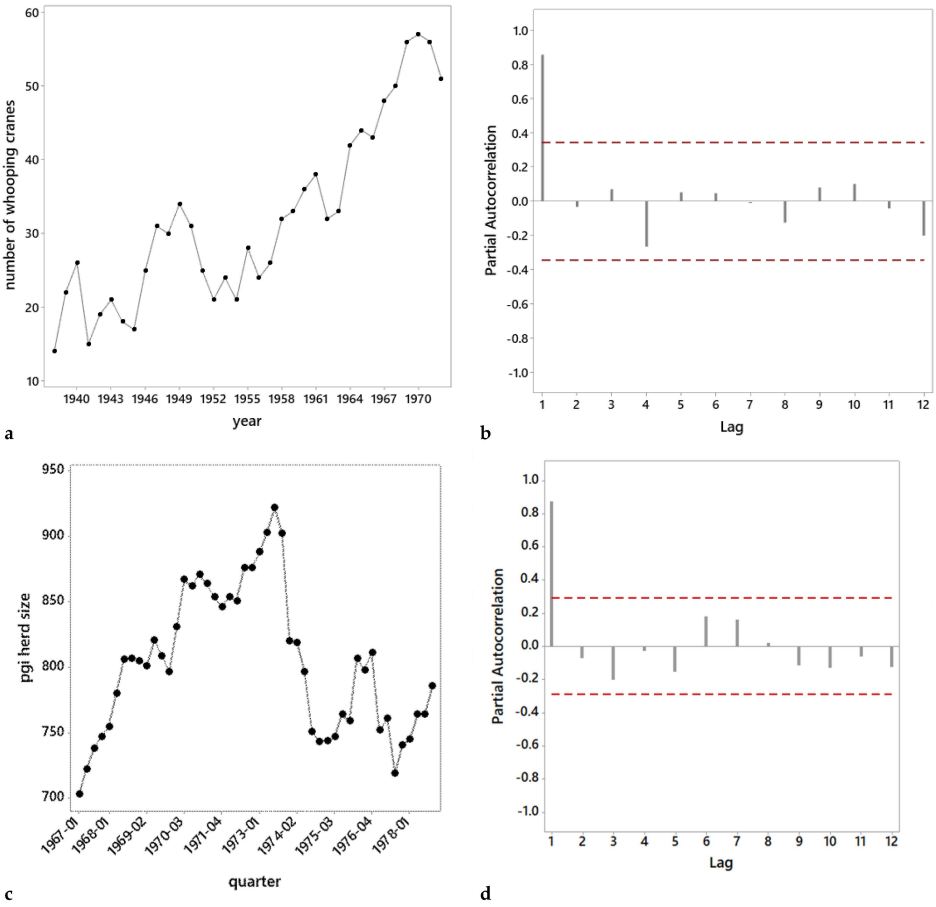

Figure 2. Time series and partial temporal autocorrelation correlogram (with 95% confidence intervals denoted by red dashed lines) plots for selected publicly available datasets. Top left (**a**): a count of whooping cranes arriving in Texas each autumn, from Hand et al. [19], Table 121. Top right (**b**): Figure 2a correlogram. Bottom left (**c**): United Kingdom pig herd size data, from Andrews and Herzberg [17], Table 62.1. Bottom right (**d**): Figure 2c correlogram.

Today, numerous time series datasets exist, some dating back several decades. Both Andrews and Herzberg [17] and Hand et al. [19] include annual lynx trapping counts (respectively, Tables 3.1 and 109), and monthly sunspots (respectively, Tables 11.1 and 112). Andrews and Herzberg [17] also include annual lynx pelt counts and prices (Table 3.2), body temperature measurements (Tables 48.1 and 48.4), monthly ozone thickness measures (Table 12.1), the Earth's annual rotation change angles (Table 20.1), hourly, daily, and weekly rainfall quantities (Tables 13.2, 14.1, and 15.1), and monthly employment (Tables 65.1–65.4) and unemployment (Table 64.1) figures. Meanwhile, Hand et al. [19] also include weekly and monthly sales figures (Tables 245 and 107), monthly airline passenger counts (Table 113), university enrollment figures (Table 116), daily rainfall quantities (Table 157), monthly temperature measures (Table 341), monthly lung cancer death counts (Table 326), and various sequences of time intervals (Tables 160, 234, and 255).

In summary, the main point of this section is a reinforcement of the notion that time series constitute correlated data.

3.3. Correlated Data: Spatial Autocorrelation

Analogous to time series, spatial series also constitute a category of correlated data [22]; Cressie [23] furnishes an excellent overview of the common models used to describe this variety of data. A spatial data series does not necessarily have a discernable pattern in its observations' covariance matrix (see Reference [24], reprinted in Hand et al. [19], Table 270; [25]). If a landscape is linear, then this covariance matrix includes:

$$\begin{pmatrix} 1 & -\rho & 0 & & 0 & 0 \\ -\rho & 1 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1 & & 0 & 0 \\ & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & -\rho \\ & & & & -\rho & 1 \end{pmatrix} = \mathbf{I}_n - \rho \begin{pmatrix} 0 & 1 & 0 & & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 \\ & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ & & & & 1 & 0 \end{pmatrix}, \quad (16)$$

where $\mathbf{I}_n$ is an n-by-n identity matrix, and the right-hand matrix is the spatial weights matrix $\mathbf{C}_s$; this matrix has $(n-1)(n-2)$ zeroes, and is reminiscent of the time series covariance matrix structure. A P-by-Q (i.e., $n = PQ$ for a complete rectangular grid with P rows and Q columns of polygons) regular square tessellation, such as that associated with a remotely sensed or other digital image whose data form a pixel mesh (i.e., raster data), has as its basic structure matrix:

$$\mathbf{C}_s = \mathbf{I}_P \otimes C_Q + \mathbf{C}_P \otimes \mathbf{I}_Q, \quad (17)$$

where a right-hand side matrix subscript denotes the dimension of the square matrix to which it is attached; matrices $\mathbf{C}_Q$ and $\mathbf{C}_P$ are of the preceding linear case form [26]. For this specification, the number of zeroes is $n^2 - 5n + 2(P + Q)$. Meanwhile, for a planar surface partitioned into n mutually exclusive and collectively exhaustive polygons, the maximum number of zeroes is $n^2 - 13n + 24$; its density has a minimum of $200/n\%$ and a maximum of $600(n-2)/[n(n-1)]\%$, with its calculation given by $100\mathbf{1}^T\mathbf{C}_s\mathbf{1}/n\%$.

The preceding time series discussion reveals that the popular description of temporal autocorrelation involves a second-order covariance matrix for observations. This also is the case for spatial series (e.g., the simultaneous autoregressive [27] and autoregressive response specifications). Frequently the basic correlation structure matrix for geospatial data is row-standardized (i.e., a rescaling such that each row sum of its cell entries, which all are non-negative, is one), converting it from matrix $\mathbf{C}_s$ to matrix $\mathbf{W}_s$. One intuitively appealing feature of this modification is that the maximum positive spatial autocorrelation parameter ρ value always is 1. Furthermore, for the examples presented in this section, the definition of geographic neighbor is an areal unit adjacency based upon a shared non-zero length polygon boundary (i.e., the rook definition, utilizing an analogy with chess piece moves). In general, because the operating correlation mechanisms are two/three-dimensional and multi-directional, socio-economic/demographic georeferenced data tend to contain a moderate degree of positive spatial autocorrelation described by $0.4 < \rho < 0.6$, whereas remotely sensed georeferenced data tend to contain substantial positive spatial autocorrelation described by $0.90 < \rho < 0.95$; in contrast, many time series contain autocorrelation values in the interval (0.95, 1.00) (e.g., the average daily temperature in Honolulu for a 58-year period [28]).

Andrews and Herzberg [17] furnish Table 67.1, a spatial series (based upon zip code zone areal units) dataset for Chicago insurance provision (Figure 3a, which is based upon their map); zip code zones without data as well as two zip code zones creating an isolated island in the northeastern part of the city are excluded from this illustrative analysis, whereas one included zip code zone constitutes two geographically separated areal units ($n = 44$, $\hat{\rho} = 0.68$, density = 8.5%). For this example, the sample

variance (without adjusting for covariates) estimate decreases from 1.4^2 (ignoring observational correlation) to 0.8^2 (accounting for observational correlation), and n* = 8. Meanwhile, most meaningful remotely sensed images have at least hundreds-of-thousands, if not millions, of pixels. The following are several available relatively small (but substantively rather meaningless) specimen images often used for illustrative purposes: Getis and Ord (16-by-16, with a single remotely sensed image variable [29]); High Peak Landsat 7 (30-by-30, with seven spectral bands [30]); and, Houston pre- and post-Hurricane Harvey paired pixel NDVI differences from Landsat 8 (41-by-41, Griffith, Chun, and Li, https://personal.utdallas.edu/~{}yxc070300/sra_esf/ or https://github.com/ywchun/sp_esf; also see https://giscrack.com/list-of-spectral-indices-for-sentinel-and-landsat/). The example presented here employs this last dataset because the two smaller ones require spatial covariance matrices of an order higher than two (presumably because of amplified edge/boundary effects attributable to small sample sizes). The spatial series (pixel areal units) data for the difference between pre- and post-Harvey (6 April 2017, and 3 January 2018) NDVI values appears in Figure 3c (n = 1681, $\hat{\rho}$ = 0.93, density = 0.2%). For this example, n* = 48.

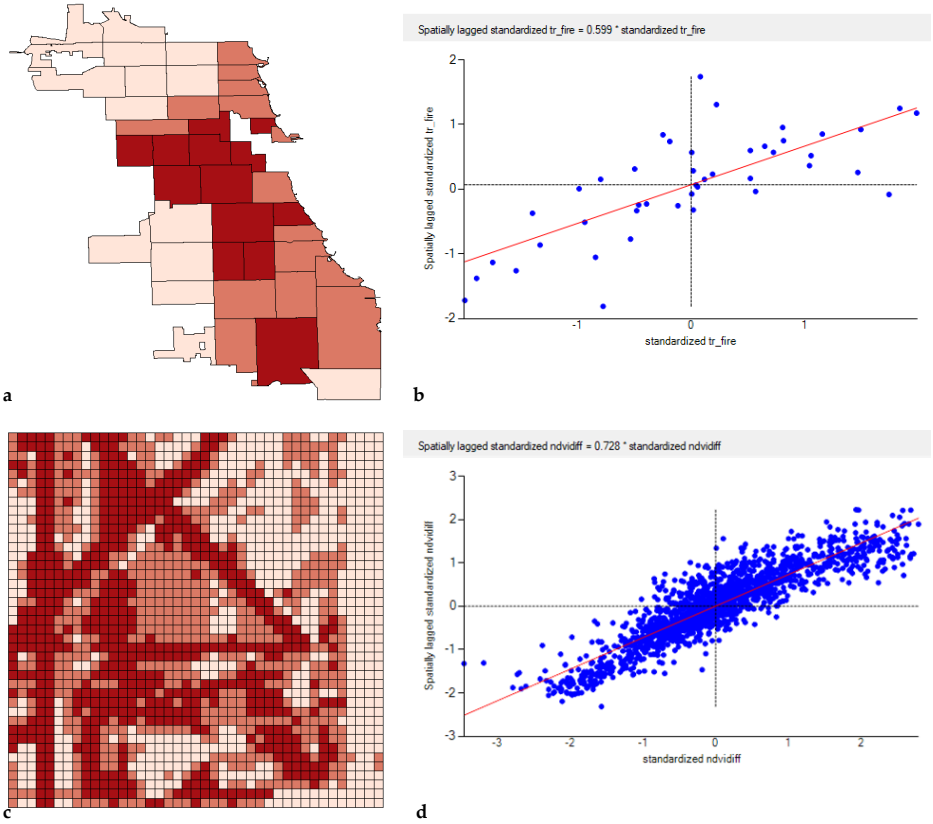

Figure 3. Selected publicly available spatial series dataset visualizations. Tertile (light brown denotes relatively small values, medium brown denotes intermediate values, and dark brown denotes relatively large values) map portrayals of spatial series and their corresponding Moran scatterplots. Top left (**a**): Box-Cox transformed fire insurance rates from Andrews and Herzberg [17], Table 67.1. Top right (**b**): A Moran scatterplot for Figure 3a. Bottom left (**c**): NDVI differences between pre- and post-Harvey (this hurricane struck during 26–30 August 2017) for a specimen region of the Houston metropolitan area. Bottom right (**d**): A Moran scatterplot for Figure 3c.

Today, numerous spatial series datasets exist, mostly because of the widespread popularity and dissemination of geographic information systems (i.e., GISs); a map needs to accompany each of these series for their adequate descriptions, which occurs only for Tables 5.1, 6.1–6.2, 16.1, 18.1, 49.1, 52.1, and 67.1 in Andrews and Herzberg [17], with Hand et al. ([19], p. ix) arguing that their deliberate

exclusion of maps maintains dataset presentation simplicity. Both Andrews and Herzberg [17] and Hand et al. [19] include agricultural field plot yields (respectively, Tables 5.1 and 320). Andrews and Herzberg [17] also include additional agricultural field plot yields (Tables 6.1 and 58.1–58.2), groundwater chemicals and soil assay findings (Tables 16.1, 17.1, and 18.1), locational accuracy repeated measures (Table 10.1)—a dataset also relating to Sections 3.1 and 3.4—species distribution counts (Table 49.1), and earthworm biomass density by constructed quadrats (Table 52.1). Meanwhile, Hand et al. [19] additionally include United States city and state crime statistics (Tables 134, 262, and 356), yeast counts and sand sedge presence/absence by constructed quadrats (Tables 163, 260), pine trees by stands (Table 250), temperature measures by cities (Table 262), industrial employment figures by countries (Table 363), and village dialect similarities (Table 145).

In summary, the main point of this section is a reinforcement of the notion that space series constitute correlated data.

3.4. Correlated Data: Space-Time Autocorrelation

Combining space and time series expands the size of an observations covariance matrix to nT-by-nT, where n denotes the number of geographic areal units, and T denotes the number of time periods; Cressie and Wikle [31] furnish an excellent overview of the common models used to describe this variety of data. In this context, the simplest version of an inverse covariance matrix includes a factor that may be written in terms of the following two distinctive specifications:

$$\text{lagged}: \ \mathbf{I}_{nT} - \rho_T \mathbf{C}_T \otimes (\rho_s \mathbf{C}_s \ + \ \mathbf{I}_n) \ = \ \mathbf{I}_{nT} \ - \ \mathbf{C}_T \otimes (\rho_T \rho_s \mathbf{C}_s \ + \ \rho_T \mathbf{I}_n), \tag{18}$$

and contemporaneous:

$$I_{nT} \ - \ (\rho_s I_T \otimes C_s \ + \ \rho_T C_T \otimes I_n), \tag{19}$$

where $\mathbf{I}_{nT} = \mathbf{I}_n \otimes \mathbf{I}_T$ is an nT-by-nT identity matrix, matrices $\mathbf{C}_s$ and $\mathbf{C}_T$ respectively are n-by-n and T-by-T and denote the spatial and temporal observations linkage structures (i.e., equivalent to those appearing in expressions (9) and (10)), and ρ_s and ρ_T respectively denote the spatial and temporal autocorrelation parameters. The lagged description furnished by expression (11) states: what occurs at location i at time t is a function of what occurred at location i at time t − 1, as well as what occurred at the neighbors of location i at time t − 1; geographic observation correlation requires time to materialize as inertia accumulates in data. The contemporaneous description furnished by expression (12) states: what occurs at location i at time t is a function of what occurred at location i at time t − 1, as well as what occurs at the neighbors of location i at time t; geographic observation correlation almost instantaneously materializes as inertia accumulates in data. Matrix expressions (11) and (12) tend to have a higher density than a pure time series, but a lower density than a pure space series, observations covariance matrix; the calculation corresponding to expression (11) is given by $100[T\mathbf{1}^T\mathbf{C}_s\mathbf{1} + n(T-1)]/[nT(nT-1)]\%$, with the sum of geographic neighbors term $\mathbf{1}^T\mathbf{C}_s\mathbf{1}$ typically not having a closed-form expression.

By the late 1980s and early 1990s, numerous time series and spatial series datasets had become readily available. In contrast, relatively few space-time datasets had become accessible, although this situation changed considerably by the dawn of the new millennium. Andrews and Herzberg [17] and Hand et al. [19] highlight this former paucity of space-time datasets: the former scholars include two (i.e., Tables 5.1 and 50.2), whereas the latter scholars include none, in their books. One noticeable characteristic of most space-time datasets is that temporal autocorrelation dominates correlation among observations, being much more pronounced than spatial autocorrelation because of its one-direction single linkage accumulation of inertia in a system. However, Andrew and Herzberg's Table 5.1, a tabulation of 74 years (1852–1924) of agricultural wheat and straw yields for a linear arrangement of 17 experimental field plots (the Broadbalk field at Rothamstead Experimental Station), fails to display this feature, most likely because conducting annual randomized crop experiments would tend to prevent an accumulation of inertia in a time series (Grondona and Cressie [32] note that one goal

of experimental design randomization involving areal units is to neutralize spatial autocorrelation; sequential years' randomizations also tend to neutralize temporal autocorrelation). The space-time inverse covariance structure matrix here has a density of 11.8%. The 17 wheat and straw time series produce sets of $\hat{\rho}_T$ values spanning the respective intervals (−0.50, 0.62) and (−0.63, 0.61); these two quite wide ranges defy a simplifying assumption of a constant ρ_T value for this pair of space-time series. Meanwhile, the 74 wheat and straw space series produce sets of $\hat{\rho}_s$ values spanning the respective intervals (−0.26, 0.33) and (−0.19, 0.44); these two rather wide ranges also defy a simplifying assumption of a constant ρ_s value for this pair of space-time series, although their small sample size most likely also contributes to this degree of dispersion. Consequently, this particular space-time series dataset seems suitable for a mixed model analysis, in which a time-invariant random effects (RE) term is estimated, exploiting the repeated measures nature of the time series for the given set of areal units. This RE term functions as a common factor across the 74 time series, and can be decomposed into spatially structured (SSRE) and spatially unstructured (SURE) components. Figure 4a,b portray scatterplots visualizing the observed-predicted pairs of values based upon estimated REs. Figure 4c,d respectively visualize the SSRE and SURE components with tertile maps describing the Broadbalk field; $\hat{\rho}_s = -0.30$ (which is not significantly different from zero; its H_0 probability is 0.20).

Meanwhile, Andrew and Herzberg's Table 50.2, a tabulation of 96 months (1963–1971) of Southern Germany fox rabies cases for a 32-by-32 grid of constructed quadrats, is another space-time dataset, with Table 50.1 reporting its cumulative counts over time by these quadrats in map form. Unfortunately, counts reported for the 96 individual maps have a sum that is 13 greater than that for the collective map; furthermore, the cumulative maximum count for the former is 50, whereas it is 20 for the latter. In addition, 11 monthly quadrat maps have missing entries somewhere in their recordings; three of these maps have two, and one has four, missing entries. Consequently, this specific publicly available dataset is too corrupt to allow a proper space-time analysis of it, one that should display an accumulation of inertia. The importance of noting these mistakes here is that so many courses and textbooks, as well as databases such as that for the R project, include the Andrew and Herzberg tables for example data analyses (a contention confirmed by a cursory internet search); almost certainly, these inaccuracies compromise pedagogy. Fortunately, the United States Census Bureau furnishes numerous easily accessible space-time datasets (see https://www.census.gov/), one of which is decennial population counts for Florida counties since 1930; although some earlier data dating back to 1821 also are available, the last of the 67 contemporary Florida counties was not established until 1925.

For illustrative purposes, this section utilizes a Florida population density dataset covering nine decennial census periods, which is too few time series observations for sensibly assessing and estimating temporal autocorrelation; the simplest of analyses requires a time series with a minimum of about 50 observations [33]. Therefore, selected preliminary exploratory data analyses were undertaken using the annual estimated Florida county population counts from 1970 to 2019 (see http://edr.state.fl.us/Content/population-demographics/data/CountyPopulation_2016.pdf) that provide 67 time series with 50 observations each. The space-time inverse covariance structure matrix here has a density of 1.1%. Initial findings include that population density should be subjected to a logarithmic transformation to better mimic a bell-shaped curve, and that log-transformed population density time series should be differenced, converting the sequences from simply log-density to change in log-density time series (this differencing dramatically degrades a conformity with normality). These latter sequences display strong temporal autocorrelation described by a second-order autoregressive specification, suggesting that the lagged space-time specification would be more suitable than the contemporaneous specification. Spatial autocorrelation across the 50 maps is moderate-to-strong, with $\hat{\rho}_s$ spanning the interval (0.49, 0.69); spatial autocorrelation across the 49 differenced maps is more diverse, with $\hat{\rho}_s$ spanning the interval (−0.16, 0.66). One principal consequence of this differencing is that the illustrative space-time dataset now has 8×67, rather than 9×67, observations; employing the space-time lagged specification further reduces this number to $7 \times 67 = 469$ observations. The resulting autocorrelation parameter estimates are: $\hat{\rho}_s = 0.17$, and $\hat{\rho}_T = 0.45$; again, temporal autocorrelation overshadows

spatial autocorrelation. For this example, the sample variance (without adjusting for covariates) estimate decreases from 1.4^2 (ignoring observational correlation) to 0.2^2 (accounting for observational correlation with a random effects linear regression specification), and n* = 66 (based upon a STAR model specification).

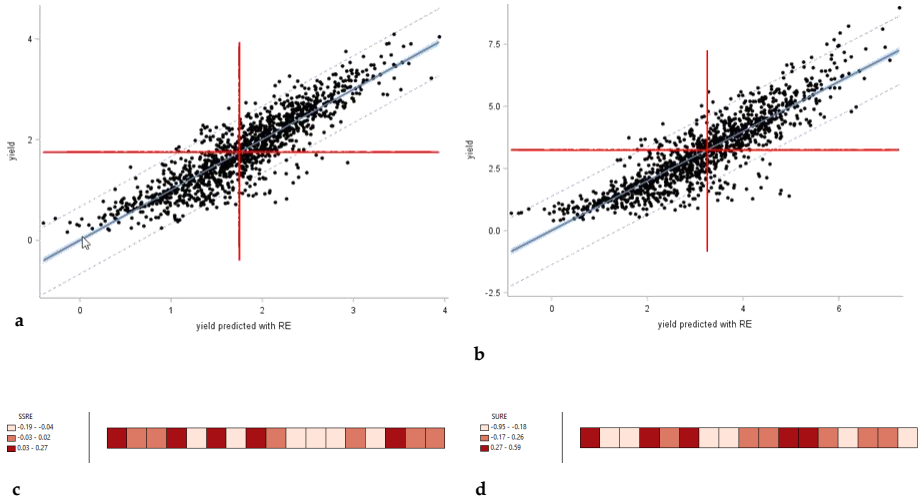

Figure 4. Selected publicly available space-time dataset visualizations. Top left (**a**): a scatterplot of the RE-based predicted and observed wheat yield values with fixed effects annual time intercepts. Top right (**b**): a scatterplot of the RE-based predicted and observed straw yield values with fixed effects annual time intercepts. Bottom left (**c**): a tertile map of the SSRE component. Bottom right (**d**): a tertile map of the SSRE component.

In summary, the main point of this section is a reinforcement of the notion that space-time series constitute correlated data.

3.5. Correlated Data: Social Network Autocorrelation

The undirected graph-theoretic representation of the preceding inter-observations correlation structures renders sparse adjacency matrices that tend to increase in density, although only rather modestly, but nevertheless constitute a small percentage of all possible adjacency matrices (e.g., see Reference [34]). For example, many spatial series relate to connected planar graphs (e.g., Reference [35]), whereas some spatial series, which employ the queen adjacency definition (another chess move analogy) or a particular set of geographic nearest neighbors or areal units within buffer zone geographic distances, relate to connected nearly-planar graphs; these low-density correlation structure articulations continue to embrace a very small percentage of all possible graphs when quantified by matrix density. More recently, correlated data analysis attention has turned to social network observations correlation structures, which relate to higher density inverse covariance structure matrices. The dataset collections compiled by Andrews and Herzberg [17] and Hand et al. [19] fail to include these types of specimen datasets because few sources for them existed prior to the dawn of contemporary network science [14]; Facebook and Flickr (launched in 2004), Twitter (launched in 2006), Instagram (launched in 2010), and Snapchat (launched in 2011), to name a few, were unknown in the 1980s and 1990s.

Hashmi et al. [36] tabulate densities for the following selected 500-node social networks: Epinions, 27.5%; Wikipedia, 23.3%; Twitter, 6.2%; e-mail, 4.8%; and, authors, 4.8%. Comparable geographic planar graphs would have densities less than 0.2%. Faust [37] summarizes descriptive statistics for 51 social networks, reporting that their densities range from 2% to 86%. In addition, the KONECT project (http://konect.cc/) makes 1326 networks available to the public. Of these, 1197 have their density (which, employing KONECT terminology, equals $100 \times$ *Fill*) tabulated; many appear to have a very

low density (186 are nearly zero), similar to the preceding correlated data categories, with the largest density being 79.5% (only 10 networks have a density of at least 50%). Nevertheless, a number of these densities are substantially greater than any of those for the preceding inter-observation correlation structure **C** matrices.

Gatewood and Price [38] analyze a social network of jazz musicians (Reference [39]; n = 198, density = 14.1%) that is available in this KONECT dataset collection. Their response variable is the degree of centrality within this network, which is the principal eigenvector of n-by-n matrix **C** (Figure 5a). This specimen network furnishes one illustrative example here. Similarly, the centrality index for a University Rovira i Virgili (Tarragona) e-mail network (Reference [40]; n = 1133, density = 0.9%) furnishes a second social network illustration (Figure 5c). These two graphs, whose Cartesian coordinates in their portrayals here are the first two eigenvectors of their doubly-centered matrix $(\mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n)\mathbf{C}(\mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n)$ appearing in the numerator of the adapted spatial autocorrelation Moran [41] coefficient index, conspicuously differ from the preceding correlated data structure graphs. Not surprisingly, the existing spatial statistical techniques, which originally evolved from time series techniques, can be exploited to analyze social network autocorrelated data (e.g., the Moran scatterplot [42]; Figure 5b,d). For the jazz musicians social network, $\hat{\rho} = 0.82$, and $n^* = 7$, whereas for the university e-mail social network, $\hat{\rho} = 0.46$, and $n^* = 317$. Respectively, after a conventional standardizing of the two principal eigenvectors (i.e., dividing each element by its vector's maximum, and then multiplying by 100), the sample variances (without adjusting for covariates) estimate decrease from 20.9^2 and 10.4^2 (ignoring observational correlation) to 15.7^2 and 9.1^2 (accounting for observational correlation).

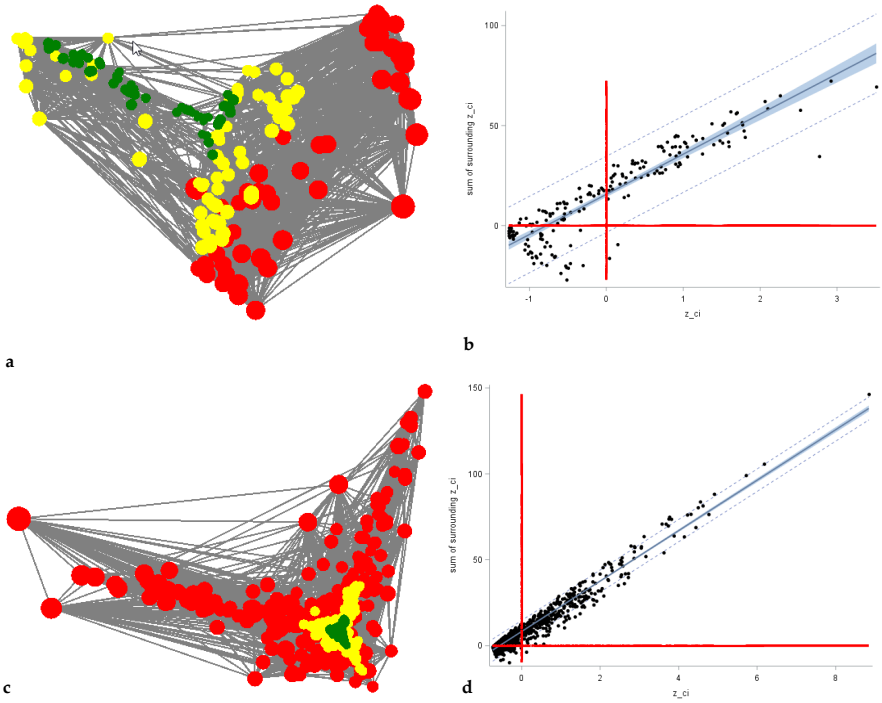

Figure 5. Selected publicly available social network series dataset visualizations. A topological centrality index portrayed by choropleth proportional circles; green denotes relatively small values, yellow denotes intermediate values, and red denotes relatively large values. Top left (**a**): a jazz musician social network. Top right (**b**): A Moran scatterplot for Figure 5a, with 95% confidence (solid blue) and prediction (dotted blue lines) intervals. Bottom left (**c**): an e-mail social network. Bottom right (**d**): A Moran scatterplot for Figure 5c, with 95% confidence (solid blue) and prediction (dotted blue lines) intervals.

In summary, the main point of this section is confirmation of the notion that social network series constitute correlated data. In doing so, it sets the stage for social network autocorrelation being a fundamental item for big data discussions.

4. Discussion and Implications

The initial empirical example appearing in each section documents one well-known consequence of ignoring existing observational correlation in real-world data, namely, variance inflation. As the repeated measures example illustrates, this inflation can produce nontrivial changes in such test statistics as t and F. In other words, it compromises inferential bases for statistical decision making.

Meanwhile, numerical results summarized in the preceding discussion support a simple but informative comparison between the Liu-Liang and the proposed new classification schemes. On the one hand, the preceding triplets of Pythagorean quadruples can produce the following illuminating results based upon the Liu-Liang categories:

Assumption	**Independent**	**Exchangeable**	**Autoregressive ($\hat{\rho}_T$ = 0.48)**	**Unstructured**
pairwise correlations	0, 0, 0	0.40, 0.40, 0.40	0.48, 0.23, 0.48	0.53, 0.22, 0.44

Equicorrelation and heteroscedastic variance results for the exchangeable category are nearly identical here because of the variance stabilizing transformations employed. A rudimentary outcome is that the Liu-Liang classification fails to furnish an effective differentiation between the simpler case of matched observations (with its equicorrelation going to zero as the number of matches increases) and the other cases of inter-observations correlation, rendering virtually the same autoregressive and unstructured results. In other words, one view of the Liu-Liang scheme is that assumptions govern it. On the other hand, reported social network series results for n = 500 enable the following comparison across categories of the proposed scheme:

Assumption	**In-Dependent**	**Matched Pairs**	**Time Series**	**Space Series**	**Space-Time Series**	**Social Network: Epinions**
matrix density	0%	0.2%	0.2%	1.1%	1.2%	27.5%

The space series results are an arithmetic average of 100 simulated planar surface partitionings [35] of size n = 500 based upon the rook adjacency definition (the average number of neighbors for this sample ranges from 4.2 to 5.7, consistent with corresponding averages for empirical surface partitionings). Likewise, the space-time series results are an arithmetic average of 100 simulated planar surface partitionings of size n = 100 based upon the rook adjacency definition (the average number of neighbors for this sample ranges from 4.3 to 5.5) coupled with five consecutive time periods for a contemporaneous space-time covariance structure specification. A rudimentary outcome here is that the proposed classification scheme does a much better job of differentiating among its categories; in other words, a matrix density evidence-based taxonomy appears to be superior to an assumption-based one. A main point of this section is the highlighting of benefits offered by this new taxonomy.

One extension meriting attention that the discussion in Section 3.4 alludes to is a fuller explication of the time-invariant RE specification for space-time data. The Broadbalk example involves crop yields that randomization prevented from accumulating inertia over time and space. In contrast, the Florida change in population density demonstrates this type of inertia accumulation. Figure 6a,b respectively portray the predictive capabilities of the space-time autoregressive (STAR) and the RE specifications for this Florida case. One considerable advantage of the RE specification is that, unlike the space-time

lagged specification, it does not sacrifice a year of data for estimation purposes (i.e., nT = 536). Both specifications account for roughly 30% of the variation in the log-population density change over time. The RE term's two components, the SSRE and the SURE (respectively, Figure 6c,d), each account for roughly half of this RE percentage of variation. In addition, with this specification, $\hat{\rho}_s = 0.66$, which is noticeably greater than its counterpart for the space-time autoregressive specification. Figure 6c primarily depicts a contrast between southern Florida and much of the Florida panhandle.

A principal implication of the correlated observations classification scheme proposed in this paper is that another category of inter-observations correlation structure remains to be discovered, one that most frequently would have densities in the 75% to 95%+ range; presumably 100% density cases always will be rare, a hypothesis meriting future investigation. Smith [43] argues that this 100% density context poses new mathematical statistics difficulties and complications: spatial statistical models with extremely high densities produce downward biased maximum likelihood estimates of the spatial dependence parameter ρ (see Section 3.3). This downward bias also pertains to tests for spatial autocorrelation. One serious consequence is the possibility that positive spatial autocorrelation may be present but undetected. Smith ([43], p. 324) notes that " . . . the actual severity of these biases can be determined only by [additional,] more extensive and systematic simulation studies," or other more comprehensive numerical and mathematical statistical investigations. For this completely connected case, with 100% density,

$$n* = -\frac{1.10356}{1 + e^{6.00290}} + \frac{1.10356}{1 + e^{6.00290 - 8.24725\sqrt{\rho}}} + (1-\rho)^2 n, \tag{20}$$

where this first term ranges from 0 to nearly 1 as positive ρ increases; again, n* = n for ρ = 0, and n* = 1 for ρ = 1.

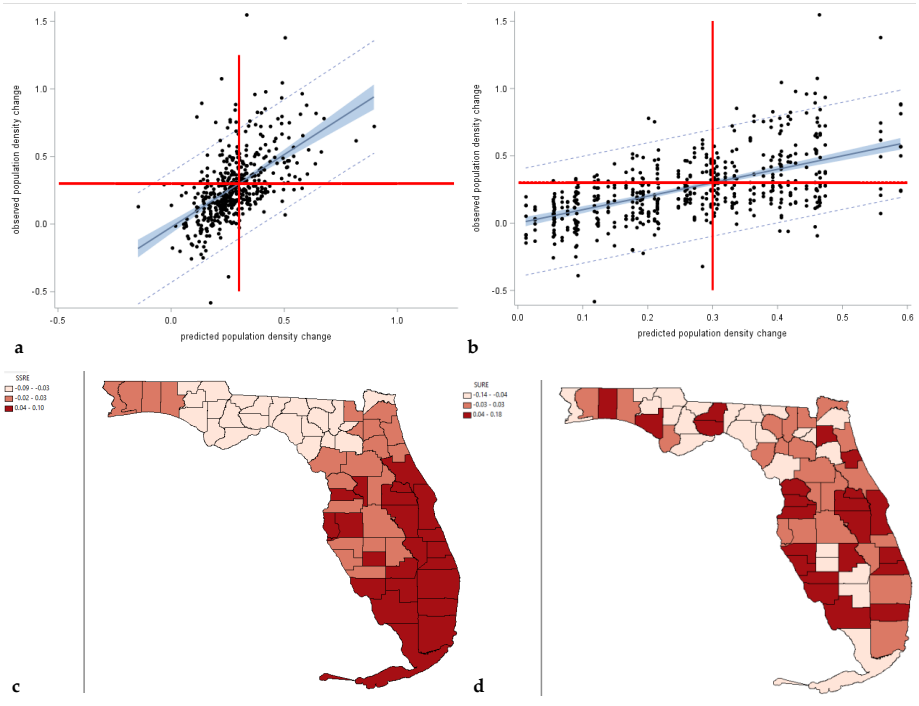

Figure 6. Selected publicly available space-time dataset visualizations of change in Florida decennial census log-population density, 1930-2010. Top left (**a**): a scatterplot of the space-time autoregressive predicted and observed values, with 95% confidence (solid blue) and prediction (dotted blue lines) intervals. Top right (**b**): a scatterplot of the RE-based predicted and observed values, with 95% confidence (solid blue) and prediction (dotted blue lines) intervals. Bottom left (**c**): a tertile map of the SSRE component. Bottom right (**d**): a tertile map of the SURE component.

Several authors already note and are undertaking needed relevant future research efforts. Liang and Zeger [44], for example, discuss the use of regression with correlated observations (e.g., Section 3.1), whereas Xia [45], for example, examines extensions of the regression tree paradigm to clustered binary outcomes. Hanley et al. [46] argue that techniques for handling correlated observations, such as generalized estimating equations, need to be made more accessible to the applied statistics community; this claim certainly also can be made about an understanding of the inverse covariance correlation structure matrices that are the topic of this paper. Finally, Zhang and Wu [47], for example, address the generalization of goodness-of-fit tests to correlated data situations. Meanwhile, the contents of this paper suggest additional needed future pursuits. One is the derivation of an effective sample size formula based upon the RE specification. Another is establishing the appropriate inverse covariance matrix density for an RE specification. A third is exploring what happens when $-1 < \rho < 0$, the case of negative autocorrelation (e.g., Reference [48]). Finally, emphasizing publicly available datasets in this paper reveals more than bothersome compilation/recording/reporting errors in these data sources, data corruptions that compromise needed reproducibility and replicability (e.g., Reference [49]) for data analytic advances. This situation calls for an accessible archive of correct specimen datasets that covers the spectrum of data types emphasized by the new classification scheme proposed in this paper. Another main point of this section is the urgency and seriousness of this situation: because of their rarity in the past, datasets like Andrew and Herzberg's Southern Germany fox rabies dominate the sample datasets analyzed in statistics courses and textbooks, and appear in such collections as the R datasets.

5. Conclusions

In conclusion, this paper presents a new family of correlated data, one based upon sparsity of the inverse of a set of observations' covariance matrix; more specifically, the accompanying structure matrix formulation $(\mathbf{I} - \rho\mathbf{C})$, which may be of this exact first-order correlation form (e.g., matched observations), or the popular second-order correlation form $(\mathbf{I} - \rho\mathbf{C})^{T}(\mathbf{I} - \rho\mathbf{C})$ (e.g., spatial, temporal, and network autocorrelation). One important issue here pertains to the estimation of covariance matrix density for the different classes of models, a rather unchallenging task for modern computers and moderate-size datasets when employing the simplest of specifications presented in this paper for illustrative purposes; subsequent research needs to address these classes of correlated data when matrix parametrizations employ a larger number of parameters (e.g., semivariogram models for space, time, and space-time data defining the denser covariance matrices rather than their inverses) and when densities are nearly 100%. Graph-theoretic density of the inverse covariance structure matrix differentiates among members of this family, beginning with 0% off-diagonal entries for independent observations, and ending with as much as approximately 86% off-diagonal entries for social network autocorrelation. Calculation of an effective sample size also is feasible for these members, with n not changing for independent observations, and n reducing to n/k with k repeated measures (for which n is a function of k), and nearly one for other extreme high inter-observations correlation cases, regardless of matrix density. The category of exchangeable correlation structures (a la the de Finetti probability theorem) forming part of an alternative scheme outlined by Liu and Liang [9] can occur in various density-based correlation structures. In addition, the Liu-Liang unstructured category can occur with repeated measures, the lowest density correlated data member. Closing comments in the preceding section address comparisons between the two classification families, as well as extensions, implications, and ongoing further developments for descriptions and properties of the proposed family presented in this paper.

Funding: This research received no external funding.

Acknowledgments: The author is an Ashbel Smith Professor of Geospatial Information Sciences.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Todem, D. Statistical analysis of longitudinal and correlated data. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin, Germany, 2011; pp. 1383–1386.
2. Held, L.; Schwab, S. Improving the reproducibility of science. *Significance* **2020**, *17*, 10–11. [CrossRef]
3. Sainani, J. The importance of accounting for correlated observations. *Phys. Med. Rehabil.* **2010**, *2*, 858–861. [CrossRef]
4. Miller, J. Earliest Known Uses of Some of the Words of Mathematics. Available online: http://jeff560.tripod.com/mathword.html (accessed on 28 June 2020).
5. Baker, A. The early history of average values and implications for education. *J. Stat. Educ.* **2003**, *11*, 1. [CrossRef]
6. David, H. First (?) occurrence of common terms in mathematical statistics. *Am. Stat.* **1995**, *49*, 121–133.
7. Fisher, R. The Correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **1918**, *52*, 399–433. [CrossRef]
8. Legler, J.; Roback, P. Broadening Your Statistical Horizons. 2019. Available online: https://bookdown.org/roback/bookdown-bysh/ (accessed on 22 February 2020).
9. Liu, G.; Liang, K.-Y. Sample size calculations for studies with correlated observations. *Biometrics* **1997**, *53*, 937–947. [CrossRef] [PubMed]
10. Stigler, S. *The History of Statistics: The Measurement of Uncertainty before 1900*; Harvard University Press: Cambridge, MA, USA, 1986.
11. Griffith, D. Spatial statistics: A quantitative geographer's perspective. *Spat. Stat.* **2012**, *1*, 3–15. [CrossRef]
12. Hotelling, H. The generalization of Student's ratio. *Ann. Math. Stat.* **1931**, *2*, 360–378. [CrossRef]
13. Keller, G. The theoretical relation between scintillation and shadow bands. *Astron. J.* **1954**, *59*, 326. [CrossRef]
14. Barabási, A.-L. *Network Science*; Cambridge University Press: New York, NY, USA, 2018.
15. Stein, S. Sums and products of jointly distributed random variables: A simplified approach. *J. Stat. Educ.* **2005**, *13*. [CrossRef]
16. Muñoz, A.; Carey, V.; Schouten, J.; Segal, M.; Rosner, B. A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* **1992**, *48*, 733–742. [CrossRef] [PubMed]
17. Andrews, D.; Herzberg, A. *Data: A Collection of Problems from Many Fields for the Student and Research Worker*; Springer: New York, NY, USA, 1985.
18. Bouchard, T.; McGue, M. Familial studies of intelligence: A review. *Science* **1981**, *212*, 1055–1059. [CrossRef] [PubMed]
19. Hand, D.; Daly, F.; Lunn, A.; McConway, K.; Ostrowski, E. *A Handbook of Small Data Sets*; Chapman & Hall: New York, NY, USA, 1994.
20. Hürlimann, W. Exact and asymptotic evaluation of the number of distinct primitive cuboids. *J. Integer Seq.* **2015**, *18*, 1–9.
21. Blakeley, B.; Gal, D.; Gelman, A.; Robert, C.; Tackett, J. Abandon statistical significance. *Am. Stat.* **2019**, *73*, 235–245.
22. Ives, A.; Zhu, J. Statistics for correlated data: Phylogenies, space, and time. *Ecol. Appl.* **2006**, *16*, 20–32. [CrossRef]
23. Cressie, N. *Statistics for Spatial Data*; Wiley: New York, NY, USA, 1993.
24. Cressie, N. Geostatistics. *Am. Stat.* **1989**, *43*, 197–202.
25. Hodges, J.; Reich, B. Adding spatially-correlated errors can mess up the fixed effect you love. *Am. Stat.* **2010**, *64*, 325–334. [CrossRef]
26. Gasim, A. First-order autoregressive models: A method for obtaining eigenvalues for weighting matrices. *J. Stat. Plan. Inference* **1988**, *18*, 391–398. [CrossRef]
27. Ord, K. Estimation methods for models of spatial interaction. *J. Am. Stat. Assoc.* **1975**, *70*, 120–126. [CrossRef]
28. Palmer, H. Annual march of daily mean temperatures at Honolulu. *Pac. Sci.* **1950**, *4*, 50–54.
29. Getis, A.; Ord, J. Spatial Analysis: Modelling in a GIS Environment. In *Local Spatial Statistics: An Overview*; Longley, P., Batty, M., Eds.; Geoinformation International: Cambridge, UK, 1996; pp. 261–277.
30. Bailey, T.; Gatrell, A. *Interactive Spatial Data Analysis*; Longman: Clemsford, UK, 1995.
31. Cressie, N.; Wikle, C. *Statistics for Spatio-Temporal Data*; Wiley: New York, NY, USA, 2011.

32. Grondona, M.; Cressie, N. Using spatial considerations in the analysis of experiments. *Technometrics* **1991**, *33*, 381–392. [CrossRef]
33. Hanke, J.; Wichern, D. *Business Forecasting*, 9th ed.; Pearson: Upper Saddle River, NJ, USA, 2013.
34. Read, R.; Wilson, R. *An Atlas of Graphs*; Oxford University Press: New York, NY, USA, 2005.
35. Griffith, D. Generating random connected planar graphs. *GeoInformatica* **2018**, *22*, 767–782. [CrossRef]
36. Hashmi, A.; Zaidi, F.; Sallaberry, A.; Mehmood, T. Are all social networks structurally similar? In *A Comparative Study Using Network Statistics and Metrics*; IEEE: Piscataway, NJ, USA, 2014.
37. Faust, K. Comparing social networks: Size, density, and local structure. *Metodološki Zvezki* **2006**, *3*, 185–216.
38. Gatewood, J.; Wood, C. Utilizing social network analysis to study communities of women in conflict zones. *J. Humanist. Math.* **2017**, *7*, 3–21. [CrossRef]
39. Arenas, A. Jazz Musicians Network Data. 2009. Available online: http://deim.urv.cat/~{}alexandre.arenas/data/welcome.htm (accessed on 28 June 2020).
40. Arenas, A. E-mail Network URV Data. 2009. Available online: http://deim.urv.cat/~{}alexandre.arenas/data/welcome.htm (accessed on 28 June 2020).
41. Moran, P. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [CrossRef] [PubMed]
42. Anselin, L. The Moran Scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial Analytical Perspectives on GIS Fischer*; Scholten, M.H., Unwin, D., Eds.; Taylor and Francis: London, UK, 1996; pp. 111–125.
43. Smith, T. Estimation bias in spatial models with strongly connected weight matrices. *Geogr. Anal.* **2009**, *41*, 307–332. [CrossRef]
44. Liang, K.-Y.; Zeger, S. Regression analysis for correlated data. *Annu. Rev. Public Heal.* **1993**, *14*, 43–68. [CrossRef]
45. Xia, R. Statistical Issues in the Analysis of Correlated Data. Ph.D. Thesis, University of Michigan, Ann Arbor, MI, USA, 2015, unpublished doctoral dissertation.
46. Hanley, J.; Negassa, A.; Edwardes, M.; Forrester, J. Statistical analysis of correlated data using generalized estimating equations: An orientation. *Pract. Epidemiol.* **2003**, *157*, 364–375. [CrossRef]
47. Zhang, H.; Wu, Z. Generalized Goodness-of-Fit. Tests for Correlated Data. *arXiv* **2018**, arXiv:1806.03668. Available online: https://arxiv.org/abs/1806.03668v1 (accessed on 28 June 2020).
48. Griffith, D. Negative spatial autocorrelation: One of the most neglected concepts in spatial statistics. *Stats* **2019**, *2*, 27. [CrossRef]
49. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*; The National Academies Press: Washington, DC, USA, 2019.