32

Smoothed ANOVA Modeling

Miguel A. Martinez-Beneito
Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO)
Valencia, Spain
and
CIBER de Epidemiología y Salud Publica (CIBERESP)
Madrid, Spain

James S. Hodges
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, Minnesota

Marc Marí-Dell'Olmo
CIBER de Epidemiología y Salud Pública (CIBERESP)
Madrid, Spain
and
Agencia de Salud Pública de Barcelona
Barcelona, Spain
and
Institut de Investigació Biomédica (IIB Sant Pau)
Barcelona, Spain

CONTENTS

Smoothed analysis of variance, usually known as SANOVA, was proposed in different forms with different goals by Nobile and Green [1], Gelman [2], and Hodges et al. [3]. This chapter builds on the latter, which proposed a method for smoothing effects in balanced ANOVAs having a single error term, that is, without random effects as understood by, for example, Scheffé [4]. Zhang et al. [5] applied this approach to multivariate disease mapping as a

T&F Cat #K23899 — K23899_C032 — page 585 — 12/10/2015 — 17:32

simpler alternative to the intrinsic multivariate conditional autoregressive (MCAR) distribution, often used to analyze multivariate areal data (Section 1 of Zhang et al. [5] gives citations to pertinent MCAR literature). This application of SANOVA used specific known linear combinations of the diseases under study, presumably with particular meanings, to structure the covariance among diseases, which in most multivariate analyses is usually assumed to be unknown and unstructured [6, 7]. More recently Marí-Dell'Olmo et al. [8] proposed a reformulation of SANOVA for disease mapping that is simpler to implement and allows extensions such as multivariate ecological regression and spatiotemporal modeling.

This chapter reviews the SANOVA approach and shows some modeling possibilities it allows. The chapter is organized as follows: Section 32.1 introduces the original formulation of SANOVA for multivariate disease mapping and the advantageous reformulation. Section 32.2 discusses some settings where this approach can be applied, beyond its original use for multivariate modeling. Finally, Section 32.3 shows a multivariate ecological regression of mortality data in Barcelona, Spain, illustrating one use of SANOVA and the powerful epidemiological conclusions that can be drawn from it.

32.1 Smoothed ANOVA

For now, we consider the following multivariate disease mapping problem. Let O_{ij} and E_{ij} denote, respectively, the number of observed and expected health events in the ith geographical unit ($i = 1, \ldots, I$) for the jth outcome under study ($j = 1, \ldots, J$). From now on, without loss of generality, we refer to counties when talking of areal geographical units and to diseases when talking about outcomes. We assume

$$O_{ij} \sim Poisson(E_{ij} \exp(\mu_{ij})).$$

The multivariate disease mapping problem is mostly concerned with how to model $\boldsymbol{\mu}$, the matrix of log standardized mortality ratios (SMRs), to represent dependence both within diseases (spatial dependence) and between diseases.

32.1.1 Zhang et al.'s SANOVA proposal

SANOVA for multivariate disease mapping was proposed by Zhang et al. [5], using as an example the incidence of $J = 3$ cancers in the $I = 87$ counties of Minnesota. The idea was to model $vec(\boldsymbol{\mu}) = (\boldsymbol{\mu}'_{\cdot\mathbf{1}}, \ldots, \boldsymbol{\mu}'_{\cdot\boldsymbol{J}})'$, where each $\boldsymbol{\mu}'_{\cdot\boldsymbol{j}}$ is an I-vector, using a two-way ANOVA without replication, with factors disease and county. Because the number of diseases is usually much smaller than the number of counties, the disease main effect was modeled as a set of fixed effects. The proposed model did not include one fixed effect (indicator variable) for each disease, but rather one fixed effect for each of J specified linear combinations of the diseases. The coefficients of those linear combinations were arranged as the columns of a matrix $\boldsymbol{H}$. Zhang et al. proposed to set $\boldsymbol{H}_{\cdot\mathbf{1}}$ to $J^{-1/2}\mathbf{1}_{\boldsymbol{J}}$, so the first linear combination corresponds to the ANOVA's grand mean. The remaining columns of $\boldsymbol{H}$ were called $\boldsymbol{H}^{(-)}$, so $\boldsymbol{H}$ may be written as $[\boldsymbol{H}_{\cdot\mathbf{1}} : \boldsymbol{H}^{(-)}]$. $\boldsymbol{H}^{(-)}$ was specified so that $(\boldsymbol{H})'\boldsymbol{H} = \boldsymbol{I}_{\boldsymbol{J-1}}$; that is, the columns of $\boldsymbol{H}^{(-)}$ are orthogonal contrasts describing specific features of the diseases. Obviously, $\boldsymbol{H}^{(-)}$ could be defined infinitely many ways, yielding different SANOVA models. The choice of a specific $\boldsymbol{H}^{(-)}$ would depend on the questions of interest to the modeler. This is similar to a traditional ANOVA, in which the selection of a specific set of contrasts usually depends on the questions to be answered or the statistical design used to answer them.

Thus, the disease main effect's contribution to the model for $vec(\boldsymbol{\mu})$ has this form:

$$\left(\boldsymbol{H} \otimes (I^{-1/2}\mathbf{1}_{\boldsymbol{I}})\right)\boldsymbol{\Theta}_{\boldsymbol{Dis}} = \left(\boldsymbol{H}_{\cdot\mathbf{1}} \otimes (I^{-1/2}\mathbf{1}_{\boldsymbol{I}})\right)\Theta_{GM} + \left(\boldsymbol{H}^{(-)} \otimes (I^{-1/2}\mathbf{1}_{\boldsymbol{I}})\right)\boldsymbol{\Theta}_{\boldsymbol{Contrast}}, \tag{32.1}$$

where Θ_{GM} denotes the first component of $\boldsymbol{\Theta}_{\boldsymbol{Dis}}$, used to model the grand mean, and $\boldsymbol{\Theta}_{\boldsymbol{Contrast}}$ is a $(J-1)$-vector modeling the effects of the contrasts. The vector $I^{-1/2}\mathbf{1}_{\boldsymbol{I}}$ applies the J disease effects to each of the I spatial regions of $vec(\boldsymbol{\mu})$; $I^{-1/2}$ is a normalizing constant.

Conversely, the number of counties is usually large in this kind of setting, which precludes modeling them as fixed effects. Moreover, it is convenient to use the counties' geographical arrangement to define dependence among their respective risks, especially given that counties are small areas. Thus, counties were modeled as a set of spatially correlated random effects. Zhang et al. proposed an intrinsic CAR distribution for modeling counties, with precision matrix $\tau\boldsymbol{Q}$, where $Q_{ii} = m_i$, the number of county i's neighbors, and $Q_{ii'} = -1$ if counties i and i' are neighbors and 0 otherwise. Let $\boldsymbol{Q}$ have spectral decomposition $\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'$, where $\boldsymbol{V}$ is an orthogonal matrix and $\boldsymbol{D}$ is diagonal. In the sequel, we assume the region of study defines a connected map (i.e., it consists of a single connected island), so $\boldsymbol{D}$ has exactly one diagonal element equal to 0 [9], which we assume to be the first diagonal element, contrary to the usual convention of sorting $\boldsymbol{D}$'s diagonal elements in decreasing order. Note that the eigenvector corresponding to that zero eigenvalue is $\boldsymbol{V}_{\cdot\mathbf{1}} = I^{-1/2}\mathbf{1}_{\boldsymbol{I}}$. We denote as $\boldsymbol{V}^{(-)}$ the $I \times (I-1)$ submatrix of $\boldsymbol{V}$ containing the columns with nonzero diagonal elements in $\boldsymbol{D}$, so $\boldsymbol{V}$ may be written as $[\boldsymbol{V}_{\cdot\mathbf{1}} : \boldsymbol{V}^{(-)}]$. Similarly, $\boldsymbol{D}^{(-)}$ denotes the submatrix of $\boldsymbol{D}$ with the first row and column removed. Zhang et al. proposed to model the county main effect as $\boldsymbol{V}^{(-)}\boldsymbol{\Theta}_{\boldsymbol{County}}$, where $\boldsymbol{\Theta}_{\boldsymbol{County}} \sim N_{I-1}(\mathbf{0}, (\tau\boldsymbol{D}^{(-)})^{-1})$, which yields the precision matrix

$$\left(\boldsymbol{V}^{(-)}(\tau\boldsymbol{D}^{(-)})^{-1}(\boldsymbol{V}^{(-)})'\right)^{-1} = \tau(\boldsymbol{V}^{(-)}\boldsymbol{D}^{(-)}(\boldsymbol{V}^{(-)})') = \tau\boldsymbol{Q}.$$

This model is equivalent to an intrinsic CAR distribution on the county main effect, which is a random effect (though not in the sense used by, e.g., Scheffé [4]). The contribution of the county random effect to the model for $vec(\boldsymbol{\mu})$ therefore has the following form:

$$\left(J^{-1/2}\mathbf{1}_{\boldsymbol{J}} \otimes \boldsymbol{V}^{(-)}\right)\boldsymbol{\Theta}_{\boldsymbol{County}} = \left(\boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{V}^{(-)}\right)\boldsymbol{\Theta}_{\boldsymbol{County}}, \tag{32.2}$$

where the term $J^{-1/2}\mathbf{1}_{\boldsymbol{J}}$ applies the I county effects to each of the J diseases considered.

If the model included no more effects, the risks of all J diseases would have the same geographical pattern except for differences in their intercepts arising from the disease main effect. An interaction between disease and county is needed to allow deviation from this additive structure. The design matrices of the disease and county effects in Equations 32.1 and 32.2 are built using the components of the matrix modeling the between-disease structure $\boldsymbol{H} = [\boldsymbol{H}_{\cdot\mathbf{1}} : \boldsymbol{H}^{(-)}]$ and the components of the matrix modeling the spatial structure $\boldsymbol{V} = [\boldsymbol{V}_{\cdot\mathbf{1}} : \boldsymbol{V}^{(-)}]$. Thus, the design matrix for the grand mean is just $\boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{V}_{\cdot\mathbf{1}}$; for the contrasts in the columns of $\boldsymbol{H}^{(-)}$, that is, the disease main effect, the design matrix is $\boldsymbol{H}^{(-)} \otimes \boldsymbol{V}_{\cdot\mathbf{1}}$; and for the county main effect, the design matrix is $\boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{V}^{(-)}$. It seems natural therefore for the disease-by-county interaction to have design matrix $\boldsymbol{H}^{(-)} \otimes \boldsymbol{V}^{(-)}$, combining the dependence between diseases defined by $\boldsymbol{H}^{(-)}$ with the spatial dependence structure in $\boldsymbol{V}^{(-)}$. Thus, if $\boldsymbol{\Theta}_{\boldsymbol{inter}} \sim N_{(I-1)(J-1)}(\mathbf{0}, diag(\tau_1, \ldots, \tau_{J-1}) \otimes \boldsymbol{D}^{(-)})$ and the

disease–county interaction is defined as $(\boldsymbol{H}^{(-)} \otimes \boldsymbol{V}^{(-)})\boldsymbol{\Theta}_{\boldsymbol{inter}}$, this version of SANOVA models the log SMRs as

$$\begin{aligned} vec(\boldsymbol{\mu}) &= (\boldsymbol{H} \otimes \boldsymbol{V})\boldsymbol{\Theta} = ([\boldsymbol{H}_{\cdot\mathbf{1}} : \boldsymbol{H}^{(-)}] \otimes [\boldsymbol{V}_{\cdot\mathbf{1}} : \boldsymbol{V}^{(-)}])(\Theta_{GM}, \boldsymbol{\Theta}'_{\boldsymbol{County}}, \boldsymbol{\Theta}'_{\boldsymbol{Contrast}}, \boldsymbol{\Theta}'_{\boldsymbol{Inter}})' \\ &= [\boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{V}_{\cdot\mathbf{1}} : \boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{V}^{(-)} : \boldsymbol{H}^{(-)} \otimes \boldsymbol{V}_{\cdot\mathbf{1}} : \boldsymbol{H}^{(-)} \otimes \boldsymbol{V}^{(-)}] \\ &\quad \times (\Theta_{GM}, \boldsymbol{\Theta}'_{\boldsymbol{County}}, \boldsymbol{\Theta}'_{\boldsymbol{Contrast}}, \boldsymbol{\Theta}'_{\boldsymbol{Inter}})' \\ &= (\boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{V}_{\cdot\mathbf{1}})\Theta_{GM} + (\boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{V}^{(-)})\boldsymbol{\Theta}_{\boldsymbol{County}} + (\boldsymbol{H}^{(-)} \otimes \boldsymbol{V}_{\cdot\mathbf{1}})\boldsymbol{\Theta}_{\boldsymbol{Contrast}} \\ &\quad + (\boldsymbol{H}^{(-)} \otimes \boldsymbol{V}^{(-)})\boldsymbol{\Theta}_{\boldsymbol{Inter}}. \end{aligned} \tag{32.3}$$

This model implies $vec(\boldsymbol{\mu})$ has precision matrix $\boldsymbol{Q} \otimes (\boldsymbol{H} diag(\tau, \tau_1, \ldots, \tau_{J-1})\boldsymbol{H}')$, with known $\boldsymbol{H}$ [5]. By contrast, the multivariate intrinsic CAR (MCAR) distribution has a precision matrix of the form $\boldsymbol{Q} \otimes \boldsymbol{\Omega}$ for an unknown symmetric, positive definite $\boldsymbol{\Omega}$, the between-disease precision matrix. The fixed, known contrasts of SANOVA's $\boldsymbol{H}$ play the role of the eigenvectors of the MCAR's $\boldsymbol{\Omega}$, and the more they resemble $\boldsymbol{\Omega}$'s true eigenvectors, the better will be the fit of SANOVA. The drawback is that it is very difficult to have prior intuition about the eigenvectors of $\boldsymbol{\Omega}$ to help in specifying $\boldsymbol{H}$, although Zhang et al. presented a modest simulation experiment suggesting that in practice, this creates little or no disadvantage, most likely because the data provide weak information about $\boldsymbol{\Omega}$'s eigenvectors. Zhang et al. viewed this as a weakness of the proposed model, but Marí-Dell'Olmo et al. [8] saw it as an opportunity: if $\boldsymbol{H}$'s columns are chosen to focus on substantive questions of interest to the modeler, SANOVA becomes a way to simplify multivariate modeling of several diseases. From this viewpoint, SANOVA-based smoothing uses just J parameters $(\tau, \tau_1, \ldots, \tau_{J-1})$ to define the multivariate dependence between diseases, in contrast to MCAR, which uses $\boldsymbol{\Omega}$'s $J(J+1)/2$ parameters. In this sense, SANOVA can be considered a simpler and more convenient way to induce multivariate dependence between diseases.

32.1.2 Marí-Dell'Olmo et al.'s SANOVA proposal

The starting point of Marí-Dell'Olmo et al.'s [8] proposal is Equation 32.3. There, the log SMRs are modeled as the product $(\boldsymbol{H} \otimes \boldsymbol{V})\boldsymbol{\Theta}$, which can be expressed as

$$vec(\boldsymbol{\mu}) = (\boldsymbol{H} \otimes \boldsymbol{V})\boldsymbol{\Theta} = (\boldsymbol{H} \otimes \boldsymbol{I_I})(\boldsymbol{I_J} \otimes \boldsymbol{V})\boldsymbol{\Theta} = (\boldsymbol{H} \otimes \boldsymbol{I_I})vec(\boldsymbol{\Psi}), \tag{32.4}$$

where the random effects in the $(I \cdot J)$-vector $vec(\boldsymbol{\Psi}) = (\boldsymbol{I_J} \otimes \boldsymbol{V})\boldsymbol{\Theta}$ follow an intrinsic CAR distribution. If $\boldsymbol{\Psi} = (\boldsymbol{\Psi}'_{\cdot\mathbf{1}}, \ldots, \boldsymbol{\Psi}'_{\cdot\boldsymbol{J}})'$ for I-vectors $\boldsymbol{\Psi}'_{\cdot j}$, then Equation 32.4 can be written as

$$\begin{aligned} (\boldsymbol{H} \otimes \boldsymbol{I_I})vec(\boldsymbol{\Psi}) &= \begin{pmatrix} H_{11}\boldsymbol{I_I} & \cdots & H_{1J}\boldsymbol{I_I} \\ \vdots & \ddots & \vdots \\ H_{J1}\boldsymbol{I_I} & \cdots & H_{JJ}\boldsymbol{I_I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi}_{\cdot\mathbf{1}} \\ \vdots \\ \boldsymbol{\Psi}_{\cdot\boldsymbol{J}} \end{pmatrix} = \begin{pmatrix} H_{11}\boldsymbol{\Psi}_{\cdot\mathbf{1}} + \ldots + H_{1J}\boldsymbol{\Psi}_{\cdot\boldsymbol{J}} \\ \vdots \\ H_{J1}\boldsymbol{\Psi}_{\cdot\mathbf{1}} + \ldots + H_{JJ}\boldsymbol{\Psi}_{\cdot\boldsymbol{J}} \end{pmatrix} \\ &= \boldsymbol{H}_{\cdot\mathbf{1}} \otimes \boldsymbol{\Psi}_{\cdot\mathbf{1}} + \ldots + \boldsymbol{H}_{\cdot\boldsymbol{J}} \otimes \boldsymbol{\Psi}_{\cdot\boldsymbol{J}}. \end{aligned} \tag{32.5}$$

Therefore, Zhang et al.'s proposal can be seen as the sum of J Kronecker products of disease contrasts and the spatial patterns. Because $\boldsymbol{H}_{\cdot\mathbf{1}}$ is simply $J^{-1/2}\mathbf{1}_{\boldsymbol{J}}$, $\boldsymbol{\Psi}_{\cdot\mathbf{1}}$ contributes to the fit exactly the same way for every disease; that is, it models the component common to all the diseases, which we previously called the county main effect. $\boldsymbol{\Psi}_{\cdot\mathbf{2}}$ contributes to the fit in one way for diseases for which the corresponding element in $\boldsymbol{H}_{\cdot\mathbf{2}}$ is positive, and in the opposite way for diseases with negative elements in $\boldsymbol{H}_{\cdot\mathbf{2}}$. In general, then, for $j = 2, \ldots, J$, $\boldsymbol{\Psi}_{\cdot\boldsymbol{j}}$ models the spatial pattern associated with the jth contrast in diseases, identifying regions where this contrast takes higher or lower values. With this reformulation, SANOVA allows exploration of each contrast in which the modeler has an interest.

This reformulation of Zhang et al.'s proposal also has computational advantages. First, Zhang et al.'s approach requires that the matrix $\boldsymbol{Q}$ in the intrinsic CAR's precision matrix has no unknown parameters, so it does not extend to other spatial distributions, such as the proper CAR distribution, for which the analogous matrix and its spectral decomposition depend on unknown parameters. In that case, if MCMC was used to sample from the posterior distribution, this would require a new spectral decomposition of $\boldsymbol{Q}$ at every MCMC iteration, which could be prohibitive. Marí-Dell'Olmo et al.'s reformulation does not have this problem because computationally, it makes little difference if $\boldsymbol{\Psi_{\cdot 1}}, \ldots, \boldsymbol{\Psi_{\cdot J}}$ follow an intrinsic CAR distribution or any other spatially structured distribution. Moreover, even the graphical modeling approach in Chapter 31 could also be implemented within the SANOVA framework just introduced in order to ascertain an appropriate geographical dependence structure for the available data.

Marí-Dell'Olmo et al.'s reformulation can be used to extend the original SANOVA formulation to nonseparable multivariate dependence structures by putting different distributions on the $\boldsymbol{\Psi_{\cdot 1}}, \ldots, \boldsymbol{\Psi_{\cdot J}}$, in which case the resulting covariance structure cannot be the Kronecker product of a disease covariance matrix and a single spatial covariance. In this sense, the reformulated SANOVA generalizes the original because it can reproduce nonseparable covariance models. Moreover, Marí-Dell'Olmo et al.'s reformulation has a second advantage: it can be implemented in standard Bayesian software such as WinBUGS, OpenBUGS, or INLA. Equation 32.5 defines a SANOVA model as the sum of several Kronecker products of predefined contrasts and vectors of spatial random effects. For the jth disease, this sum of Kronecker products is

$$\boldsymbol{\mu_{\cdot j}} = H_{j1}\boldsymbol{\Psi_{\cdot 1}} + \ldots + H_{jJ}\boldsymbol{\Psi_{\cdot J}},$$

that is, a known linear combination of the spatial random effects. This simple expression of the log SMRs for any disease avoids Kronecker products and is therefore easily implemented in the aforementioned packages.

32.2 Some Specific Applications of Smoothed ANOVA

Although Zhang et al. [5] proposed SANOVA as a tool for traditional multivariate modeling in disease mapping studies, it can be used in a wider collection of settings. The contrasts in $\boldsymbol{H}$ are defined by the modeler, and this could be seen as a drawback. But these contrasts provide room for modeling; if properly used, they permit a great variety of models. For example, although $\boldsymbol{H}$ was described above as representing contrasts among levels of a single factor (diseases), with no change to the preceding theory, $\boldsymbol{H}$ can represent contrasts defining a balanced design with any number of factors, for example, a three-factor design with factors diseases, sex, and time periods. With this in mind, we now describe some settings where SANOVA can be applied for purposes somewhat different from its original conception.

32.2.1 Design-based studies in disease mapping

From their beginning, disease mapping studies have had mainly an observational aim, that is, obtaining reasonably reliable estimates for small areas to describe the geographical pattern underlying some diseases. At most, such studies may suggest the presence of a risk factor influencing the disease pattern, and this hypothesis could be tested in a confirmatory ecological regression study. Such a confirmatory study would ideally be done with new data to avoid post hoc analyses, possibly leading to the "Texas sharpshooter fallacy" [10].

Sometimes research questions involve comparing the geographical patterns of different diseases, different population groups, or different time periods. Unfortunately, traditional disease mapping methods do not address these questions; they were not conceived to do so. For example, suppose we have data for males and females for a disease and we want to explore the common geographical pattern of both sexes, as well as the geographical pattern of differences between sexes, that is, places with higher occurrence of the disease for one sex than for the other. These questions could be addressed only informally with traditional univariate disease mapping studies. Multivariate models such as MCAR include the correlation structure of the diseases and sexes, but that correlation does not necessarily address the questions of interest. Therefore, traditional disease mapping methods are not helpful. SANOVA, however, can incorporate those questions into the study's design through the matrix of contrasts $\boldsymbol{H}$. In this sense, smoothed ANOVA enables new multivariate disease mapping analyses, going beyond disease mapping's traditional descriptive purpose. Design-based studies to confirm or explore the hypothesis of interest become possible; indeed, this may require changing the descriptive conception of most disease mapping professionals.

32.2.2 Variance decomposition

The design matrices arising from the SANOVA approaches outlined in the previous section are orthonormal. Thus, if we use $J-1$ contrasts in diseases or groups in a SANOVA study, in addition to the linear combination modeling the grand mean, the design matrix's orthonormality allows us to decompose the variance of $vec(\boldsymbol{\mu})$ into these J components [8]. This decomposition can be a valuable epidemiological component in this kind of study, allowing us to see which elements of the decomposition explain most (or least) of the variance in the original data patterns.

This variance decomposition could be used, for example, in the study of lung cancer mortality in two periods and both sexes. In that case, we would define $\mathbf{H}$ with four columns: the common geographical pattern underlying all four maps (i.e., one map for each of the two periods $\times$ two sexes), the geographical pattern of differences between mortality in the two periods, the geographical pattern of differences between the mortality of the sexes, and the geographical pattern of deviations from these time and sex main effects, that is, the interaction of period and sex. If the four original sex-by-period maps are labeled so that the first two maps correspond to the first period and the first and third maps correspond to males, the $\mathbf{H}$ matrix arising from this design would be

$$\mathbf{H} = \frac{\mathbf{1}}{\mathbf{2}} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}. \tag{32.6}$$

Several obvious epidemiological questions arise here. Which of the four geographical patterns explains the most variance? Are geographical differences between sexes more important than those between periods, in terms of the variance explained? Is the sex-by-period interaction—the change between periods in the difference between sexes—important for explaining the original data pattern, after accounting for the effects of sex and period? Answers to these questions can provide important clues about the epidemiology of lung cancer. This kind of result is clearly beyond the scope of traditional univariate and even multivariate disease mapping studies.

32.2.3 Multivariate ecological regression

One more application of SANOVA is ecological regression [8, 11]. Following Marí-Dell'Olmo et al.'s approach, the original patterns in the data can be modeled or decomposed as a function of $\boldsymbol{\Psi_{\cdot 1}}, \ldots, \boldsymbol{\Psi_{\cdot J}}$, where each of these vectors contains the geographical pattern of a contrast in a column of $\boldsymbol{H}$, or the common underlying pattern in the case of $\boldsymbol{\Psi_{\cdot 1}}$. But these vectors $\boldsymbol{\Psi_{\cdot j}}$ could themselves be modeled by means of an ecological regression linking them to covariates of interest. In that case, we could determine the relationship between the covariate and the original data patterns through the estimated relationship between the covariate and the patterns corresponding to the contrasts in $\boldsymbol{H}$'s columns. In this sense, we use the contrasts to do a multivariate ecological regression study: we do not separately model the covariates' contribution to the original data patterns; we model these contributions entirely through the contrasts. Section 32.3 discusses an example in detail.

A second use of ecological regression in this context is linked to the variance decomposition described just above. As described so far, SANOVA splits the original variance into as many components as $\boldsymbol{H}$ has columns. But in ecological regression, we split these components further, into a part explained by the covariate and a second part attributed to other, possibly unknown factors, typically modeled using a spatial random effect. If the spatial random effect is defined to be orthogonal to the covariate [12, 13], we can split the variance explained by each contrast into at least two parts, the variance explained by the covariate and the "residual" unexplained variance [8]. Besides permitting the variance decomposition, placing an orthogonality restriction on the random effect avoids so-called spatial confounding, that is, confounding between the covariate of interest and the residual spatial pattern, a common problem in ecological regression problems [12].

32.2.4 Spatiotemporal modeling

Spatiotemporal problems [14] are just a kind of multivariate study with an order relationship (i.e., time sequence) on the geographical patterns being modeled, so SANOVA can be used for spatiotemporal studies [15]. In this case, it suffices to specify the columns of $\boldsymbol{H}$ as the elements of a basis of functions used to model the time trends for the geographical units composing the region of study. If an orthogonal basis is used, the variance decomposition mentioned above is retained, allowing us to explore which elements of the basis explain more and less of the variance in the spatiotemporal dataset.

The time trend for the ith geographical unit is modeled as

$$\boldsymbol{\mu_{i\cdot}} = (\boldsymbol{H_{\cdot 1}}\psi_{i1} + \ldots + \boldsymbol{H_{\cdot J}}\psi_{iJ})',$$

where $\boldsymbol{H_{\cdot j}}$ is the jth element of the basis for functions of time, evaluated at all the time units of the period of study. Since $\boldsymbol{\Psi_{\cdot j}}$ will typically have some spatial structure for each j, the parameters defining the time trend for the geographical units will be correlated: the time trends for nearby regions will be similar because they are similar combinations of the same basis elements.

The basis functions used to model time trends can be tailored to the data at hand. If no cyclic time trend is expected, a polynomial basis could be used. But if, as often happens, a cyclic trend is present, a Fourier basis could be used and will yield a much better fit. Therefore, SANOVA provides a powerful, versatile tool for modeling spatiotemporal disease mapping datasets.

Note also that other factors, such as sex or multiple diseases, can be modeled simultaneously with time, as indicated in this section's introduction.

32.3 Multivariate Ecological Regression Study Using SANOVA

We now illustrate the potential of SANOVA using chronic obstructive pulmonary disease (COPD) and lung cancer mortality data for the city of Barcelona, Spain. These two diseases have common risk factors, mainly tobacco consumption, so it seems reasonable to do a multivariate study of them. We have mortality data for both diseases and sexes, that is, observed and expected counts for all four combinations of these two factors on Barcelona's 1491 census tracts (each with about 1000–2000 inhabitants). Since tobacco consumption can be heavily influenced by deprivation, we also have this variable for every census tract so we can control for its effect, if possible.

Given these mortality data, researchers might be interested in several epidemiological questions, such as:

- Which census tracts show more mortality for all four combinations of disease and sex (common component)?
- Which census tracts show more mortality for one of the diseases regardless of sex (disease component)?
- Which census tracts show more mortality for one of the sexes regardless of disease (sex component)?
- Given the geographical distribution of the common, disease, and sex components, does the interaction between disease and sex make an important contribution to the variance of disease incidence?
- Is it possible to quantify the variability of the factors above with respect to the total variability of all four geographical patterns?
- What part of the variability of the common, disease, and sex components can be explained by deprivation?
- What is the geographical distribution of the common, disease, and sex components that cannot be attributed to deprivation?

These epidemiological questions cannot be addressed by traditional disease mapping techniques, but they can be addressed with SANOVA, as we will illustrate. The analysis we suggest is an example of what Section 32.2 called a *design-based study*, because both the design of the data to be studied and the questions to be answered make it convenient to consider specific relationships among all four geographical patterns in the study. This goal can be achieved easily using SANOVA.

From now on, we will label as geographical patterns 1 and 2 those corresponding to COPD deaths, for men and women, respectively, and label as 3 and 4 those corresponding to lung cancer deaths, also for men and women, respectively. We use expression (32.6) as the $\boldsymbol{H}$ matrix, so $\boldsymbol{\Psi_{\cdot 1}}$ represents the common component for all four geographical patterns, having higher risks for all four disease-by-sex groups than those regions i with $\Psi_{i1} > 0$. Similarly, $\boldsymbol{\Psi_{\cdot 2}}$ represents the disease-specific component, taking values higher than 1 for regions with a ratio of COPD versus lung cancer mortality higher than that for Barcelona in aggregate. $\boldsymbol{\Psi_{\cdot 3}}$ represents the sex-specific component, taking values higher than 1 for regions with a ratio of male versus female mortality higher than that for Barcelona in aggregate. Finally, $\boldsymbol{\Psi_{\cdot 4}}$ represents the disease-by-sex interaction, taking values higher than 1 for regions with particularly high mortality for COPD in men and lung cancer in women.

All four components in matrix $\boldsymbol{\Psi}$ are modeled in the same manner, as

$$\boldsymbol{\Psi_{\cdot j}} = \mu_j \cdot \mathbf{1} + f_j(\boldsymbol{D}) + \boldsymbol{S_{\cdot j}},\ j = 1, \ldots, 4,$$

where μ_j is the intercept, modeling the mean of $\boldsymbol{\Psi_{\cdot j}}$ for the whole city. As proposed in Mar-Dell'Olmo et al. [8], the expression $f_j()$ is a step function of $\boldsymbol{D}$, the deprivation index. Values of $\boldsymbol{D}$ are split into groups at specific quantiles, and $f_j()$ assigns the same value to all census tracts in the same group. In our study, we used 40 groups to define $f_j()$ and modeled $(f_j(1), f_j(2), \ldots, f_j(40))$ with an intrinsic CAR distribution, considering consecutive quantile groups as neighbors. The vector $\boldsymbol{S_{\cdot j}}$, modeled as the usual sum of heterogeneous and intrinsic CAR random effects [16], models the residual variability in each component that cannot be explained by deprivation. We impose the condition that $\boldsymbol{S_{\cdot j}}$ sums to 0 for every group defined by quantiles of the deprivation index to guarantee orthogonality of $f_j()$ and $\boldsymbol{S_{\cdot j}}$. This has two benefits: first, the variance of the original dataset can be decomposed as a function of all the terms in the model, and second, this avoids potential spatial confounding of $f_j()$ and $\boldsymbol{S_{\cdot j}}$, which otherwise could compete to explain the same variation in the data. All computations were made using INLA [17]. Further modeling and computational details are in Marí Dell'Olmo et al. [8], which used a very similar model.

Table 32.1 shows the variance explained by each component in the model. The variance attributable to the intercept of each component has not been included in Table 32.1 because it is 0 for all components. This is because expected cases have been calculated by internal standardization for each disease–sex combination, so the mean relative risk for each combination is 1, and this term does not induce any variability in the model. Among the four components considered, the common component explains the largest proportion of variance, followed by the sex and disease main effects, in that order. The disease-by-sex interaction explains hardly any variance, so henceforth we will ignore it. As a consequence, the maps for the four sex–disease combinations will be similar because of the common component's large fraction of the total variance, while maps for the two sexes will be less similar than those for the two diseases. Finally, the effect on the map of changing sex will be the same regardless of the disease, and analogously for the effect of changing disease.

Regarding the effect of deprivation, most of the variance in the data (77.5%) is associated with this factor. Nevertheless, deprivation is not equally associated with all four components considered. For instance, deprivation explains almost 99% of the variance of the difference between maps for males and females, with negligible variance attributable to other factors. On the other hand, deprivation accounts for only 68% of the variance of the difference between diseases, with, presumably, other factors underlying the remaining differences.

Figure 32.1 shows the estimated association between deprivation and the common, disease, sex, and interaction components, $\exp(f_j(\cdot)), j = 1, \ldots, 4$, respectively. All four plots show the posterior mean and 95% posterior credible interval for the 40 deprivation groups.

TABLE 32.1
Percentage of variance explained by each component in the model

Component	Variance deprivation (%)	Variance random effect (%)	Total (%)
Common	34.5	15.2	49.7
Disease	13.2	6.2	19.4
Sex	29.2	0.3	29.5
Interaction disease–sex	0.6	0.8	1.4
Total	77.5	22.5	

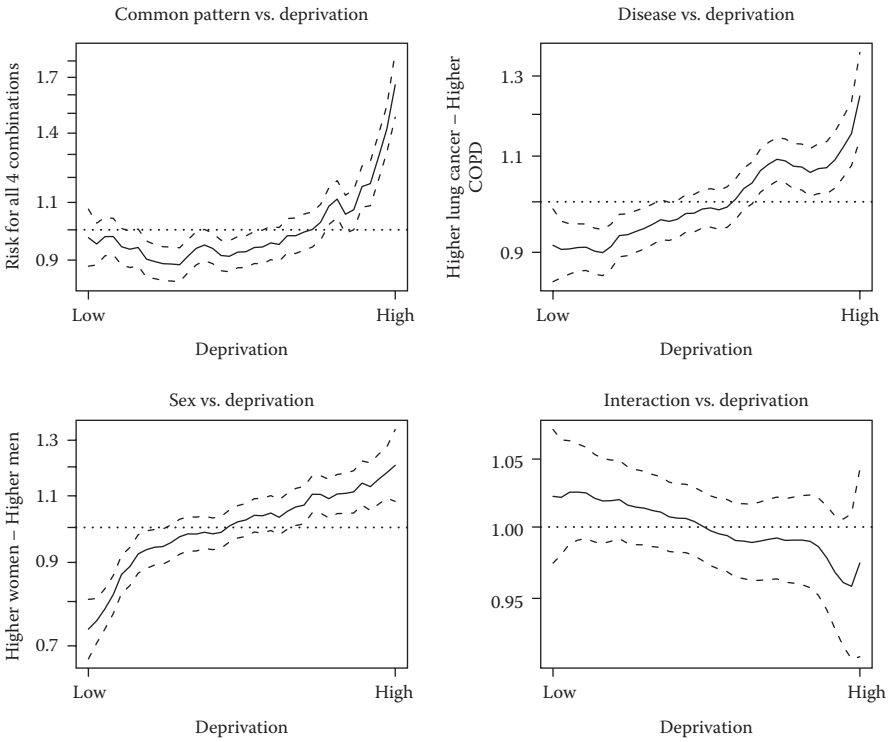

FIGURE 32.1
Relationship between deprivation and all four components included in the model.

Many deprivation levels have posterior credible intervals completely above or below 1, providing evidence that those census tracts have particular features linked to deprivation, making them different from the city's mean level. Specifically, the most deprived regions have, in general, higher mortality (for all four combinations of disease and sex), with risk up to 70% higher than the city's mean. Also, the most deprived groups show particularly high COPD mortality compared to lung cancer mortality, in contrast to the most affluent census tracts, which show the opposite trend. The most deprived regions show a higher mortality for men than for women, while the most affluent regions show higher mortality for women. This effect is especially pronounced for census tracts with the lowest deprivation, where the fitted curve has its steepest slope; this is consistent with the historically high prevalence of tobacco consumption by women in Spain's most affluent social groups [18]. As expected from Table 32.1, deprivation is not associated with the disease–sex interaction component.

Figures 32.2 and 32.3 show choropleth maps of the parts of the common and disease specific terms that are not related to deprivation, $\exp(\boldsymbol{S_{\cdot 1}})$ and $\exp(\boldsymbol{S_{\cdot 2}})$, respectively. The analogous maps for the sex and interaction components are not shown because they explain very little variance and neither has any census tract with significant excess risk compared to the city's mean risk (i.e., with 95% credible interval excluding 1). Ellipses in these figures indicate regions with significant deviations from the level of the city as a whole. Full-color versions of these figures can be found as annex material at https://www.crcpress.com/9781482253016.

The patterns in Figures 32.2 and 32.3 are uncorrelated with deprivation by construction, so they reflect the presence of other risk factors. Both components have regions with significant departures from Barcelona's mean mortality. Thus, for all four combinations of

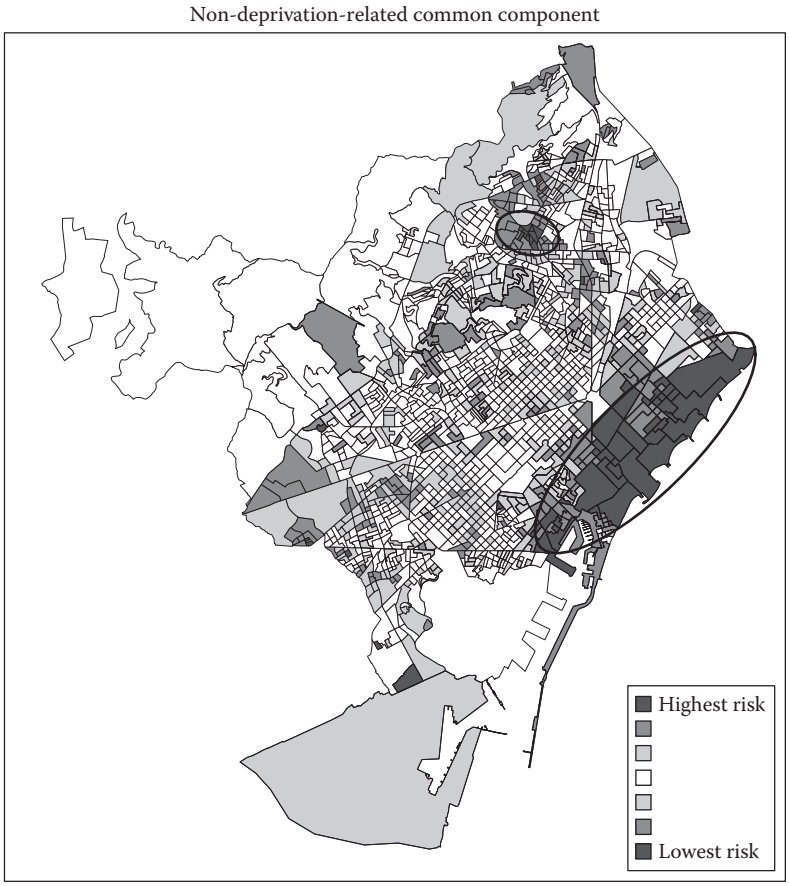

FIGURE 32.2
Geographical distribution of the part of the common component that is not related to deprivation. Ellipses indicate regions with large risk deviations compared to the whole city. Darker regions stand for larger deviations showing either higher (regions with thick border) or lower (regions with thin border) risk than the mean of the city.

disease and sex (Figure 32.2), a large region along the city's northern shoreline (the lower ellipse) has a significant risk excess. This risk excess cannot be explained by deprivation; indeed, that region includes census tracts with both the highest and lowest deprivation levels. Moreover, that region also includes both relatively new and old neighborhoods with very different demographic and social groups, suggesting an environmental risk factor as a possible explanation. Figure 32.3 shows regions with a risk excess for just one of the diseases, which cannot be explained by deprivation, pointing to the presence of risk factors for just one disease. Risk excesses have been found for both COPD (both upper-side ellipses) and lung cancer (lower-side ellipse).

This example shows that SANOVA is a powerful tool, making it possible to address questions that traditional disease mapping methods cannot. Indeed, all the questions posed at the beginning of this section have been answered using SANOVA. In this sense, SANOVA as a data analysis technique is particularly fitted to discerning mechanisms underlying diseases, beyond the exploratory aim of most disease mapping methods. This can make SANOVA a particularly appropriate tool to push disease mapping toward more analytical purposes.

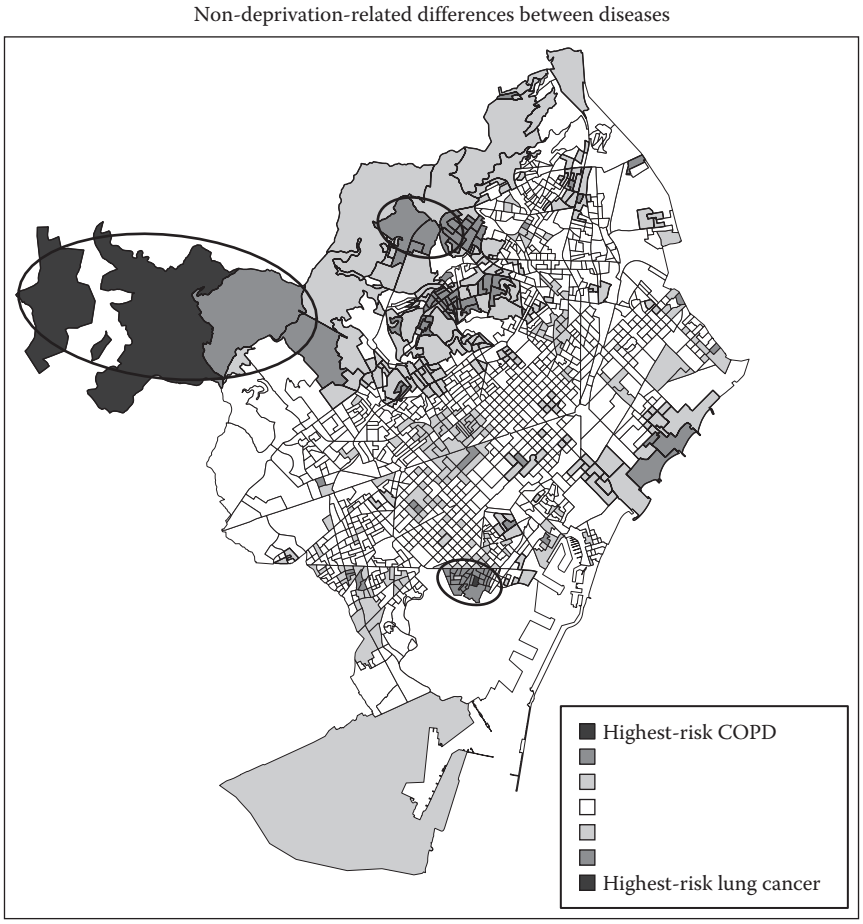

FIGURE 32.3
Geographical distribution of the part of the disease-specific component that is not related to deprivation. Ellipses indicate regions with large risk deviations compared to the whole city. Darker regions stand for larger deviations showing either higher (regions with thick border) or lower (regions with thin border) risk than the mean of the city.

References

[1] Agostino Nobile and Peter J. Green. Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, 87:15–35, 2000.

[2] Andrew Gelman. Analysis of variance—why it is more important than ever [with discussion]. *Annals of Statistics*, 33:1–53, 2005.

[3] James S. Hodges, Yue Cui, Daniel J. Sargent, and Bradley P. Carlin. Smoothing balanced single-error-term analysis of variance. *Technometrics*, 49:12–25, 2007.

[4] Henry Scheffé. *The Analysis of Variance*. Wiley, New York, 1959.

[5] Yufen Zhang, James S. Hodges, and Sudipto Banerjee. Smoothed ANOVA with spatial effects as a competitor to MCAR in multivariate spatial smoothing. *Annals of Applied Statistics*, 3(4):1805–1830, 2009.

[6] Paloma Botella-Rocamora, Miguel A. Martinez-Beneito, and Sudipto Banerjee. A unifying modeling framework for highly multivariate disease mapping. *Statistics in Medicine*, 34(9), 2015.

[7] Miguel A. Martinez-Beneito. A general modelling framework for multivariate disease mapping. *Biometrika*, 100(3):539–553, 2013.

[8] Marc Marí Dell'Olmo, Miguel A. Martinez-Beneito, Mercè Gotséns, and Laia Palència. A smoothed ANOVA model for multivariate ecological regression. *Stochastic Environmental Research and Risk Assessment*, 28(3):695–706, 2014.

[9] James S. Hodges, Bradley P. Carlin, and Qiao Fan. On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, 59:317–322, 2003.

[10] Atul Gawande. The cancer-cluster myth. *The New Yorker*, pp. 34–37, 1998.

[11] Marc Marí-Dell'Olmo, Mercè Gotséns, Carme Borrell, Miguel A. Martinez-Beneito, Laia Palència, Glòria Pérez, Lluís Cirera, et al. Trends in socioeconomic inequalities in ischemic heart disease mortality in small areas of nine Spanish cities from 1996 to 2007 using smoothed ANOVA. *Journal of Urban Health*, 91:46–61, 2014.

[12] James S. Hodges and Brian J. Reich. Adding spatially-correlated errors can mess up the fixed effect you love. *American Statistician*, 64(4):325–334, 2010.

[13] John Hughes and Murali Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B*, 75(1):139–159, 2013.

[14] Miguel A. Martinez-Beneito, Antonio López-Quílez, and Paloma Botella-Rocamora. An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27:2874–2889, 2008.

[15] Francisco Torres-Avilés and Miguel A. Martinez-Beneito. STANOVA: A smooth-ANOVA-based model for spatio-temporal disease mapping. *Stochastic Environmental Research and Risk Assessment*, 29:131–141, 2014. doi: 10.1007/s00477-014-0888-1.

[16] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathemathics*, 43:1–21, 1991.

[17] Havard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392, 2009.

[18] Anna Schiaffino, Esteve Fernandez, Carme Borrell, Esteve Salto, Montse Garcia, and Josep Maria Borras. Gender and educational differences in smoking initiation rates in Spain from 1948 to 1992. *European Jourrnal of Public Health*, 13(1):56–60, 2003.