TESE DE DOUTORAMENTO

Kriging: applying geostatistical techniques to the genetic study of complex diseases

Laura Martínez Calvo

ESCOLA DE DOUTORAMENTO INTERNACIONAL

PROGRAMA DE DOUTORAMENTO EN MEDICINA MOLECULAR

SANTIAGO DE COMPOSTELA

2020

DECLARACIÓN DA AUTORA DA TESE

KRIGING: APPLYING GEOSTATISTICAL TECHNIQUES TO THE GENETIC STUDY OF COMPLEX DISEASES

Para defensas telemáticas

D./Dna. Laura Martínez Calvo

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1) A tese abarca os resultados da elaboración do meu traballo.
2) De selo caso, na tese faise referencia ás colaboracións que tivo este traballo.
3) A tese é a versión definitiva presentada para a súa defensa e coincide coa versión enviada en formato electrónico.
4) Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.

E comprométome a presentar o exemplar impreso da tese no prazo dun mes dende que a EDIUS mo requira, así como o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En Santiago de Compostela, 1 de xuño de 2020

Asdo. Laura Martínez Calvo

AUTORIZACIÓN DOS DIRECTORES DA TESE

KRIGING: APPLYING GEOSTATISTICAL TECHNIQUES TO THE GENETIC STUDY OF COMPLEX DISEASES

D. Ángel Carracedo Álvarez
Dna. Raquel Cruz Guerrero

INFORMAN:

Que a presente tese, correspóndese co traballo realizado por Dna. Laura Martínez Calvo, baixo a nosa dirección, e autorizamos a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como directores desta non incorre nas causas de abstención establecidas na Lei 40/2015.

En Santiago de Compostela, 1 de xuño de 2020

Asdo Ángel Carracedo Álvarez

Asdo Raquel Cruz Guerrero

RESUMO

Kriging: aplicando técnicas xeostatísticas no estudo xenético de enfermidades complexas.

As enfermidades complexas presentan con frecuencia patróns de distribución xeográfica. Por iso, a integración de factores xenéticos e ambientais empregando sistemas de información xeográfica (SIX) e análises estatísticas específicas que teñan en conta a dimensión espacial dos datos resulta de grande axuda na investigación das súas interaccións xenotipo-ambiente. Os obxectivos do presente traballo foron avaliar a aplicación dunha técnica de interpolación xeostatística (*kriging*) no estudo de enfermidades complexas cunha distribución xeográfica heteroxénea e examinar o seu desempeño como unha alternativa aos métodos de imputación xenética convencionais. Empregando a esclerose múltiple como caso de estudo, o *kriging* demostrou ser unha ferramenta flexible e valiosa para a integración de información procedente de varias fontes e a distinta resolución espacial nun modelo que permite visualizar doadamente a súa distribución heteroxénea en Europa e explorar as complexas interaccións entre varios dos seus factores de risco xenéticos e ambientais xa coñecidos. Se ben o desempeño do *kriging* non mellorou os resultados obtidos coas técnicas de imputación xenética actuais, este estudo piloto puxo de manifesto o peor rendemento destas últimas para variantes raras en rexións cromosómicas con baixa densidade de marcadores.

Palabras chave: xeostatística, kriging, xenética molecular, imputación xenética, esclerose múltiple.

ABSTRACT

Kriging: applying geostatistical techniques to the genetic study of complex diseases.

Complex diseases often display geographic distribution patterns. Therefore, the integration of genetic and environmental factors using geographic information systems (GIS) and specific statistical analyses that consider the spatial dimension of data greatly assist in the research of their gene-environment interactions (GxE). The objectives of the present work were to assess the application of a geostatistical interpolation technique (kriging) in the study of complex diseases with a distinct heterogeneous geographic distribution and to test its performance as an alternative to conventional genetic imputation methods. Using multiple sclerosis as a case study, kriging demonstrated to be a flexible and valuable tool for integrating information from various sources and at a different spatial resolution into a model that easily allowed to visualize its heterogeneous geographic distribution in Europe and to explore the intertwined interactions between several known genetic and environmental risk factors. Even though the performance of kriging did not surpass the results obtained with current imputation techniques, this pilot study revealed a worse performance of the latter for rare variants in chromosomal regions with a low density of markers.

Keywords: geostatistics, kriging, molecular genetics, genetic imputation, multiple sclerosis.

TABLE OF CONTENTS

ABBREVIATIONS

1KGP: 1000 Genome Project

ACTH: adrenocorticotropic hormone

ADEM: acute disseminated encephalomyelitis

AG: alpha globulin

AIC: Akaike Information Criterion

AIM: ancestry-informative marker

ASD: autism spectrum disorder

BMI: body mass index

BNADN: Banco Nacional de ADN

CAAPA: Consortium on Asthma among African-ancestry Populations in the Americas

CADASIL: cerebral autosomal-dominant arteriopathy with subcortical infarcts and leukoencephalopathy

CD: complex disease(s)

CESGA: Fundación Centro Tecnolóxico de Supercomputación de Galicia

CIS: clinically isolated syndrome

CNS: central nervous system

CSF: cerebrospinal fluid

DIS: dissemination in space

DIT: dissemination in time

DMT: disease modifying therapy

DNA: deoxyribonucleic acid

E: East

EAE: experimental autoimmune encephalomyelitis

EBI: European Bioinformatics Institute

EBV: Epstein-Barr virus

EMA: European Medicines Agency

EP: evoked potentials

eQTL: expression Quantitative Trait Loci

ERS: environmental risk scores

ExE: environment-environment interaction

GA: glatiramer acetate

GAM: generalized additive models

GEMS: Genes and Environment in MS

GIS: Geographic Information Systems

GWAS: genome-wide association studies

GxE: genetic-environmental interaction

GxG: gene-gene interaction

HRC: Haplotype Reference Consortium

HERV: human endogenous retroviruses

HGDP: Human Genome Diversity Project

HLA: human leukocyte antigen

HPC: high-performance computing

HWE: Hardy Weinberg Equilibrium

IM: infectious mononucleosis

IQR: interquartile range

K-PCA-R: krige - Principal Component Analysis - regress model

K-R: krige - and - regress model

LAT: latitude

LD: linkage disequilibrium

LONG: longitude

MAF: minor allele frequency

MHC: Major Histocompatibility Complex

ML: machine learning

MRI: magnetic resonance imaging

MS: multiple sclerosis

MSIF: MS International Federation

N: North

N_e: effective population size (IMPUTE2)

NE: North East

NHGRI: National Human Genome Research Institute

NMOSD: neuromyelitis optica spectrum disorder

NW: North West

OCB: oligoclonal bands

OR: odds ratio

PAGE: Population Architecture using Genomics and Epidemiology

PC: principal component

PCA: principal components analysis

PPMS: primary progressive MS

PRMS: progressive-relapsing MS

PRS: polygenic risk scores

PUFA: polyunsaturated fatty acids

PVGIS: photovoltaic geographic information system

QC: quality control

RIS: radiologically isolated syndrome

RNA-seq: RNA sequencing

RRMS: relapsing-remitting MS

S: South

SD: standard deviation

SE: South East

SLE: systemic lupus erythematosus

SNP: single nucleotide polymorphism

SPMS: secondary progressive MS

SW: South West

UV-B: ultraviolet B irradiation

VD: vitamin D

VDBP: vitamin D binding protein

VDR: VD receptor

vs: versus

W: West

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

WHO: World Health Organization

1. INTRODUCTION

1.1. COMPLEX DISEASES

Complex diseases (CD) are characterized by the involvement of genetic and environmental factors. Current knowledge considers that only genetically susceptible subjects, exposed to certain environment factors will develop the disease.

As represented in Figure 1, CD are multifactorial conditions and, in most cases, the intricate network of interactions between their factors, either gene-gene (GxG), gene-environment (GxE) or environment-environment (ExE), is only partially understood.

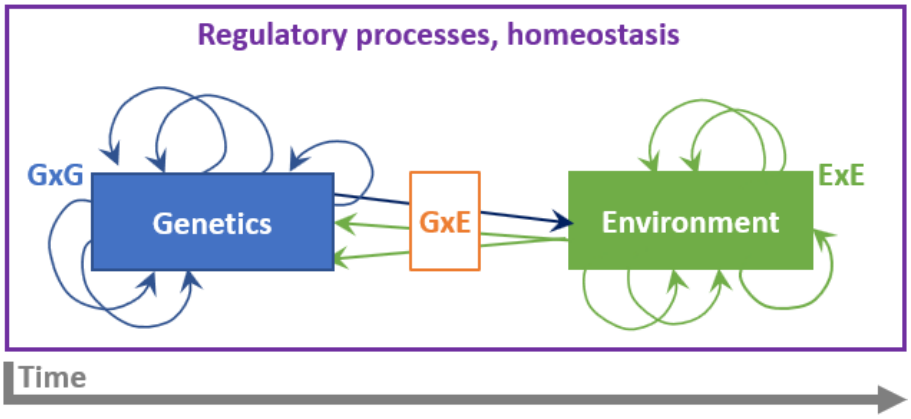

Figure 1. Schematic representation of the factors involved in the development of a complex disease.

Besides, there is a growing acceptance that both the physiopathological context and the timing in which those interactions take place (i.e. in early childhood, during a period of psychological stress) are equally important for the CD to manifest.

Due to the especial architecture of CD, the application of conventional research approaches has limitations to reach results. This

complexity claims for a holistic approach, that is, a multidisciplinary approach that allow us to have the full picture of genes, environment, and their interplay. Several projects are already collecting information in a concise manner to try to elucidate the complex relationships between factors of the disease, such as genetic and expression profiling, microbiota, biomarkers, or response to treatment data.

- UKBiobank (Sudlow et al., 2015): this project collected genotype data on ~500,000 volunteer participants as well as phenotype information (physical activity and diet questionnaires; blood, saliva and urine biomarkers; imaging data of heart and brain, among others).

- PAGE: Population Architecture using Genomics and Epidemiology (Bien et al., 2019) examined putative causal genetic variants across approximately 100,000 African Americans, Asian Americans, American Indians, European Americans, Hispanic Americans, and Native Hawaiians from four groups representing nine large U.S.-based cohorts identified in genome-wide association studies of common disease, with a large-scale effort focused on the Metabochip array, which facilitated trans-ethnic fine mapping of several outcomes, including cardiovascular diseases, stroke, obesity, diabetes or cancer.

- TOPmed: Trans-Omics for Precision Medicine (see Section 4.1.3) aims to improve scientific understanding of the fundamental biological processes that underlie heart, lung, blood, and sleep (HLBS) disorders, incorporating rich phenotypic characterization, environmental exposure data and ancestral and ethnic diversity.

Implementing this type of longitudinal studies with a full battery of screening tests and follow-ups requires a big investment, but at this point it may worth it. Even if the pathology of interest does not become fully elucidated, it may give a comprehensive picture of biological and environment processes both in health and pathological circumstances.

But what makes the study of complex diseases so challenging? The following sections will succinctly illustrate some of the points that condition their research, starting from the own disease definition and phenotype characterization to the review of the main difficulties regarding their genetic and environmental factors, as well as their complex interplay.

1.1.1 Disease definition and phenotype characterization

In most cases, the exact physiopathology of CD remains elusive: maybe some implicated biological routes are known, or a molecular pattern of the disease is already defined and used for treatment or patient stratification; although it is frequent that the initial cause of the disease is unclear. This is influenced by the fact that our knowledge of physiology under 'normal conditions' is incomplete: there are proteins without assigned function, and regulatory mechanisms implicated in diseases that we are not aware of.

Likewise, CD usually manifest through complex traits that present a range or spectrum of severity: phenotype only partially matches among affected individuals (high inter-case variation), making study's outcome definition a challenge and implying frequent misdiagnosis, undoubtedly affecting the power to detect actual effects of the disease.

1.1.2. Genetic factors[1]

The golden era of Genome-Wide Association Studies (GWAS) has enormously increased the knowledge about the biological basis of

[1]In the context of this work, the term 'genetics' will be generalized, referring to all factors related to the disease that are biologically defined and, for the most part, unmodifiable.

disease, and has triggered a huge development of methodologies, bioinformatic tools, and databases. Still, the experience has revealed some aspects with direct practical implications in their application and interpretation of results. Some of them will be briefly described next:

- Assigning clinical relevance to the findings

One of the most difficult steps in any genetic study is to infer the biologically relevant consequences of each found statistical association. Given that many found associations map to noncoding regions, in the absence of obvious functional signals, the findings of studies cannot be linked directly a biological effect.

However, continuous improvement in the understanding of the molecular machinery of gene regulation has been made in recent years, in part due to the development of methodologies that maximize the information from genetic studies playing with different levels of biological complexity (creating meta-data based on shared molecular mechanisms or physiological pathways), and predict functions or processes involved in the disease. This has already been applied in CD such as Alzheimer's disease (Habes et al., 2020), autism spectrum disorder (ASD) (Alonso-Gonzalez et al., 2019), and cancer (Fernandez-Rozadilla et al., 2019; López-Cortés et al., 2019), among many others.

- Static vs dynamic

Genetics gives us a static picture of the gene dictionary we were born with, but the level of expression varies over time to guarantee the proper response to stimuli and correct functionality, and it is cell or tissue-specific and defined to some extent by epigenetic processes.

It is thought of epigenetics as an intermediate-rate regulator of expression (not implicated in immediate response to stimuli, but modifiable in the medium-long term), initially imprinted by inheritance but changeable by the environment. This is directly related to CD known physiopathological mechanisms, being the main reason why increasing efforts are being made in order to create a comprehensive

reference epigenomics database, such as Roadmap project (Bernstein et al., 2013).

- Collected sample vs affected tissue

Over the years, blood, saliva, or bodily waste materials have been prioritized as samples due to their easy access and non-invasive characteristics, although it is widely known that the correlation between the sample and the affected tissue is not always straightforward.

Thus, correlation in expression patterns between tissues is already being studied in detail and incorporated into databases such as GTEx (The GTEx Consortium, 2013), an ongoing resource that currently has 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays such as Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), and RNA Sequencing (RNA-Seq).

- Reference population

Public genome databases, created with the effort of big consortia over the years, are intrinsically essential to genetic studies. They constitute the structural basis on with any genotyping platform would design their arrays and are some sort of encyclopedia where new findings are incorporated.

The main limitation of them is their potential sampling bias, as they are comprised of a limited number of individuals from only a few population groups. In the past decades, growing efforts have been put into obtaining different ethnic genetic backgrounds to incorporate their variation into the reference population databases (see Section 4.1.2). For example, it is frequent to consider the CEU group from the HapMap Project (Belmont et al., 2005) as a reference population for Europeans, but that sample is constituted of only 162 Utah residents with Northern and Western European ancestry.

- Reference genome

The human reference genome is an artificial construct that is periodically reviewed and updated. Consequently, databases must go

through a curation process to adapt to the newer versions of the genome assembly. Each version carries appreciable changes in the variant's chromosomal position: some variants are merged, some disappear, implying a major concern with reproducibility of results and meta-analysis, especially with single nucleotide polymorphisms (SNPs).

- Rare variation

Most existing applications for genetic studies (genotyping platforms, software for genetic association and genetic imputation; see Section 4.1) were designed for working with the main effects of common variants and tend to perform worse when applied in rare variants.

This underperformance and the increasing interest in the contribution of rare variation to disease (sometimes small but cumulative) are triggering the development of new statistical approaches and the revision of multivariate techniques for the correction of the stratification, also taking into account their great local variability.

1.1.3. Environmental factors[2]

Usually, the implication of environmental factors was solely incorporated into research as adjustment covariates, collected from epidemiologic questionnaires (sometimes patient self-reported), either retrospectively or prospectively. In the first case, there is the added constrain of completeness and reliability of data, given that some past exposures can be easily forgotten or even have taken place before birth. In the second case, a long-term follow-up is usually required, making their laborious implementation burdensome in time and cost.

[2] In the context of this work, the term 'environment' will be generalized, referring to all factors related to the disease that are highly susceptible to intervention or controlled modulation.

The main difficulty with environmental factors is not only unraveling the mechanisms by which they play their role in the disease (challenging itself as usually exposures are not recorded in a standardized way), but also the timing when they do. Sometimes, the consequences due to the role of environmental factors manifest after an accumulative process, in other times, a single exposure is enough to produce an effect.

Besides, the way an organism counteracts to environment exposure should be considered. To note, it is frequent that an (auto)immune component is involved in CD. Prove of that is that many of them share immunosuppressant treatment with effective results. Especially in late onset CD, one possible explanation is that a regulatory process that was compensating a malfunction (or the biological consequences of an exposure to environment) stops working properly (either being overtaken or being exaggeratedly active), yielding to the disease onset. What remains unclear is whether the immune system dysregulation is a cause or a consequence.

The systematic study of environmental factors involved in disease is currently having a great development. Now that the methodology for studying genetics is well-established and has proven to be reliable and successful, genetic research initiatives tend to be accompanied by the collection of information about environmental factors.

1.1.4. Interactions: GxG, ExE, GxE

As multiple forms of interplay can happen, the incorporation of interactions (either GxG, GxE or ExE) in the study of CD usually depends on how the disease is being modelled (for example, interactions are usually assumed to be additive or multiplicative, but that could not be the case (Sackton & Hartl, 2016)) and limited by the availability of data. Regarding GxE interactions, it should be noted that the sample size requirements are increased in order to ensure power (McAllister et al., 2017) and, logically, the sample size issue would

affect more when rare variants are involved, although new methodologies have been already developed to solve this (Zhao et al., 2019).

When the study of interactions has a hypothesis-generating intent, multiple machine learning (ML) approaches are gold-standard (McKinney et al., 2006; Choy et al., 2018; Habes et al., 2020), especially when data availability and computing resources are granted. However, as they are black box systems, is frequently complicated to make a direct translation of results into biologically meaningful actions.

On the other hand, when there is evidence of several genetic and environmental factors with only low to moderate size effect, one of the most widely used ways of studying their global effect is aggregating them in risk scores, weighted by their estimated size effect on the disease: this will define polygenic risk scores (PRS) and environmental risk scores (ERS), respectively. Although we get no information about the actual interaction among factors, this approach at least allows us to study their joint effect. This approach has already yielded remarkable results in CD (Xia et al., 2016; Alemany-Navarro et al., 2019; Howell et al., 2020).

As will be seen in the next section, it is worth mentioning the potential utility of a geostatistical approach in the study of CD. Considering the spatial dimension of data greatly assists in combining information from different sources.

1.2. THE GEOGRAPHIC COMPONENT OF THE DISEASE

Historically, the geographic component of disease was only subtly present. The first medical topographies appeared intending to detailly describe the environment (climate, geology, botany) and health conditions at the time. At that moment, it was thought that the variation in morbimortality could be explained by observing the social and

natural characteristics of the environment, and that it could be useful to better define public health interventions.

Over the years, the geographic perspective passed from being only a tool for visualization to consolidate as a valuable source of information itself, with full potential to identify new risk factors of the disease, detect characteristic patterns, and make future or past predictions. Among other examples, it is well known the John Snow's nineteenth-century map of cholera outbreaks in London, see Figure 2.

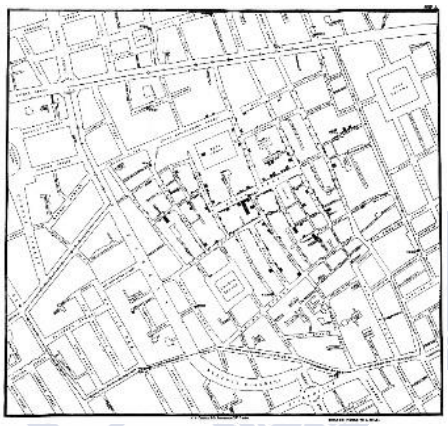

Figure 2. Original map made by John Snow in 1854. Cholera cases are highlighted in black, clustered around the pump in Broad Street.

In Genetics, geography was first incorporated to detect past natural selection processes linked to environmental factors. One of the best examples of this fact is the identification of clines[3]:

- In basal metabolism genes in indigenous circumpolar populations of North America and Siberia (Hancock et al., 2008).

- In skin pigmentation, as result of low exposure to sunlight (Jablonski & Chaplin, 2000).

[3]Cline: gradual change of a feature in a species over a geographic area; that usually goes along with a gradient in the allele frequencies of genes associated with that feature.

- In p53 expression (Shi et al., 2009; Sucheston et al., 2011), with a latitudinal gradient probably mediated by winter temperature and UV radiation.

Also, it is worth mentioning the close correlation between geography and genetic variation (Novembre & Di Rienzo, 2009; Bycroft et al., 2019), that is basic in the study of population dynamics and opens up possibilities of applying geostatistical techniques in a purely genetic context (see Section 4).

The geographic component of biological processes was established as a versatile basis on which integrate all kinds of information, defining the term **ecogeographic genetic epidemiology** (Sloan et al., 2009), illustrated in Figure 3.

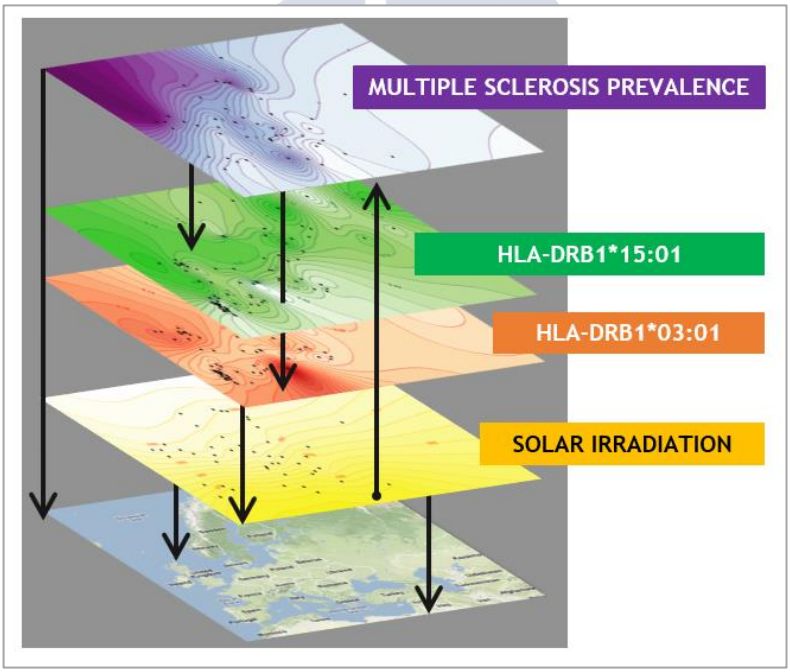

Figure 3. Representation of the integration of information from several sources through a geostatistical approach. Data from the present work.

This multidisciplinary framework set the scene for encouraging development in the biomedical sciences (Urbach & Moore, 2011), that has already been applied in different areas with successful results. To name a few: the follow-up of viral outbreaks, like the H1N1 influenza pandemic of 2009 (Jombart et al., 2011), the study of the effects of climate change in vector-borne illnesses (Colón-González et al., 2013), the mapping of the geographic distribution of infectious diseases (Dhewantara et al., 2019), the analysis of the implications of air pollution in asthma acute events (Lemke et al., 2014), the identification of spatial clusters in CD such as cancer (Ozonoff et al., 2005; Grant, 2010; Goodman et al., 2012), multiple sclerosis (Torabi et al., 2014) or rheumatoid arthritis (Källberg et al., 2013), genetic value predictions using kriging (Ober et al., 2011), prediction of complex traits with OmicKriging, using mRNA and microRNA expression data (Wheeler et al., 2014) and model-based geostatistics for disease mapping using malaria historical prevalence data (Giorgi et al., 2018).

All these advances would not be possible without the parallel development of new analysis methods or the adaptation of known methodologies (Generalized Additive Models (GAM), Bayesian geostatistical models, cluster detection…), that allow us to incorporate other forms of variation and association, considering the temporal component of data (Carrel & Emch, 2014) and that may work in an unsupervised manner (machine learning). In addition, the improvement in Geographic Information Systems (GIS) technologies (in any of their forms) permits the almost instant collection of plenty of data (Schootman et al., 2016). Digital era (along with social media) surrounds us, and there is a constant flux of information that could potentially be used for research (although the privacy and security concerns are evident). To note, the aforementioned H1N1 influenza outbreak was more reliably analyzed with data coming from Twitter than from official sources (Signorini et al., 2011).

Overall, geostatistical techniques provide a set of already implemented methods to detect and analyze geographic patterns in the distribution of diseases and identify new plausible risk or protective factors, that would definitely be helpful to have a more accurate

overview of the complex scheme that constitutes the CD. One of them (kriging) will be used in the present work. The following section reviews the main points of its application.

1.3. GEOSTATISTICS

Although this section is not intended to thoroughly cover the topic, it will informally review the basic concepts of the geostatistical technique that were applied in this work.

1.3.1. General concepts

1.3.1.1. Spatial statistics

The term spatial statistics refers to a variety of methods used for the analysis of spatially referenced data (Diggle & Ribeiro Jr, 2007). This means that spatial statistics study spatial processes in d dimensions, which can be denoted as follows

$$Z = \{Z(s), s \in D \subseteq \mathbb{R}^d\},$$

where Z is the observed attribute and D is the spatial domain (usually $D \subseteq \mathbb{R}^2$, that is, a two-dimensional domain).

Spatial data can be classified into three basic types (Cressie, 1993), defined by the characteristics of D:

- Geostatistical data or 'spatial data with continuous variation': where the domain D is a continuous, fixed set.

- Lattice data or regional data: where the domain D is fixed and discrete.

- Point data: where the domain D is random.

This work will be focused on the first type of data (geostatistical data) in a two-dimensional domain ($d = 2$).

1.3.1.2. Stationarity, isotropy, and homogeneity

The covariance function of a spatial process Z is defined as

$$Cov\big(Z(s), Z(s+h)\big)$$

Many techniques for spatial data rely upon several assumptions related with the covariance function, known as stationarity assumptions (Schabenberguer & Gotway, 2005). Second-order stationary, also known as weak stationarity, implies that the spatial process has a constant mean μ, and a covariance function which is a function of the spatial separation between points only, that is,

$$E(Z(s)) = \mu, \text{ and}$$

$$Cov\big(Z(s), Z(s+h)\big) = C(h),$$

where h is the spatial separation and function C is called the covariogram. This assumption implies the irrelevance of absolute coordinates, as the variability of the spatial process is the same everywhere if the spatial separation between two given locations is the same.

Intrinsic stationarity is weaker than second order stationarity and is commonly used in geostatistical modeling. Intrinsic stationarity is verified if the spatial process has a constant mean μ, and the variance of the difference between the process in two points depends only on the difference vector, that is,

$$E(Z(s)) = \mu, \text{ and}$$

$$Var\big(Z(s) - Z(s+h)\big) = 2\gamma(h),$$

where the functions 2γ and γ are called the variogram the semivariogram. A second-order stationary process is also intrinsically stationary, given that

$$
\begin{aligned}
&Var\big(Z(s) - Z(s+h)\big) \\
&\qquad = Var\big(Z(s)\big) + Var\big(Z(s+h)\big) \\
&\qquad - 2Cov\big(Z(s), Z(s+h)\big) \\
&\qquad = 2\left(Var\big(Z(s)\big) - Cov\big(Z(s), Z(s+h)\big)\right) \\
&\qquad = 2\big(C(0) - C(h)\big)
\end{aligned}
$$

Consequently, the semivariogram γ would be

$$\gamma(h) = C(0) - C(h)$$

However intrinsic stationary does not imply second-order stationary.

On the other hand, an intrinsic stationary process is termed isotropic if the semivariogram depends only on the absolute distance between locations $\|h\|$ (not on the direction), being $\|\cdot\|$ the Euclidean norm. An intrinsic stationary and isotropic process is called homogeneous process.

1.3.1.3. Semivariogram elements

To make predictions in an intrinsic stationary process, we will need to compute its semivariogram γ, which can be defined by three elements:

- Sill (σ^2): refers to variability in distant points; the semivariogram increases with the distance till it stabilizes when reaches a sill.

- Range: is the lag distance at which the semivariogram achieves the sill; if sill is achieved only asymptotically, the practical range is considered instead, that is, the lag distance at which the semivariogram achieves 95% of the sill.

- Nugget effect: is the magnitude of discontinuity of the origin (microscale variation with sill); it cannot be observed unless the

spacing of the sample observations is made smaller; nugget effect could also be due to measurement error or variation at small scale.

1.3.2. Spatial prediction: kriging

As any other techniques, kriging requires certain assumptions about the underlying spatial process that we intend to predict, $Z = \{Z(s), s \in D \subseteq \mathbb{R}^2\}$, and is influenced by the characteristics of the data in which it is applied (finite observed realizations of the said process). In our case, the assumptions are that observed data are spatially dependent, and underlying spatial process is intrinsically stationary and isotropic (homogeneous process). Gaussian assumption is also required for obtaining some results in the applications compiled in the next chapters, for instance, to construct confidence bands for prediction or to apply the krige-and-regress procedure (Madsen et al., 2008).

On the other hand, kriging computations need certain quantities, such as the semivariogram γ, to be known. Since this does not happen, the usual approach is the following:

- Construct an empirical variogram.

- Estimate covariance parameters by fitting a parametric model to the empirical variogram.

- Perform spatial prediction using ordinary kriging but substituting the unknown quantities by the previously estimated ones (plug-in estimation).

Next sections briefly describe each one of these steps.

1.3.2.1. Empirical variogram estimation

Knowing that the variogram can be described as the variance of the difference process, the empirical estimation of the variogram of an intrinsically stationary and isotropic process can be done easily using a simple, moment-based estimator (Matheron estimator)

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}\left(Z(s_i) - Z(s_j)\right)^2,$$

being $N(h)$ the set of locations pairs (s_i, s_j) with coordinate difference $\|s_i - s_j\| = h$, and $|N(h)|$ the number of distinct pairs in this set.

In practice, the estimation is done letting certain tolerance region around value h. The tolerance regions must be as small as possible but guaranteeing enough number of pair of points to be able to execute a stable estimation. A drawback of this way of estimating the variogram is its instability towards extreme values. That is why several robust estimates has been proposed with the introduction of a bias correction factor or the use of the median (Cressie-Hawkins robust estimator).

In this work, empirical variograms for the two considered applications were obtained using sm.variogram function of sm R package, which constructs an empirical variogram by means of a robust form of construction based on square-root absolute value differences of the data. In addition, tests for independence, isotropy and stationarity can also be computed with this function.

1.3.2.2. Parametric model fitting to empirical variogram

A valid semivariogram is conditionally negative definite, nevertheless this property is not assured for the empirical variogram. It is for this reason that, once an empirical variogram is available, a parametric model must be fitted to ensure that estimated variogram is valid, and it can be used for performing prediction.

For isotropic processes, there are several models of semivariogram to be considered: lineal, spherical, exponential, quadratic, etc. The model parameters can be estimated using different methods, for instance, the least squares method that is described below.

Let $\{Z(s_1), \ldots, Z(s_n)\}$ be a realization of the spatial process Z, and let $\gamma_\theta(u)$ be a parametric model of variogram, being θ a set of parameters that collect the spatial dependence. Let $\{u_1, \ldots, u_k\}$ a set of

distances. From the sample $\{Z(s_1), \dots, Z(s_n)\}$, a pilot estimator $\hat{\gamma}$ is obtained (empirical variogram, etc.). Then the least squares estimator of θ is computed as follows

$$\hat{\theta} = \arg\min_{\theta} \sum_{l=1}^{k} (\hat{\gamma}(u_l) - \gamma_\theta(u_l))^2$$

Analogously, the generalized least squares estimator can be constructed as

$$\hat{\theta} = \arg\min_{\theta} (\hat{\gamma} - \gamma_\theta) \Sigma_{\hat{\gamma}}^{-1} (\hat{\gamma} - \gamma_\theta),$$

where $\Sigma_{\hat{\gamma}}$ is the covariance matrix of the estimated variogram $\hat{\gamma}$ in the distances $u_1, \dots, u_k$. Likewise, the weighted least squares estimator can be computed replacing in the previous expression the covariance matrix $\Sigma_{\hat{\gamma}}$ by a diagonal matrix such that, for $l \in \{1, \dots, k\}$, each entry (l, l) is given by (Schabenberguer & Gotway, 2005):

$$2 \frac{\gamma_\theta^2(u_l)}{|N(u_l)|}.$$

The function variofit of geoR R package was used in this work to fit exponential models to empirical variograms for the two analyzed applications. This function estimates covariance parameters by fitting a parametric model to an empirical variogram by using ordinary or weighted least squares. Some relevant aspects from practical point of view related to variofit function are detailed below.

- Numerical minimization

The parameter values are found by numerical optimization using one of the following functions: optim, nlm and nls. In given circumstances the algorithm may not converge to correct parameter values when called with default options and it is necessary to pass extra options for the optimizers. For instance, the function optim takes a control argument. In addition, different initial values should be tried,

and using options to scale the parameters may be necessary when the parameters have different orders of magnitude. Some possible workarounds in case of problems include:

- Rescale data values dividing by a constant.

- Rescale coordinates subtracting values and/or dividing by constants. This approach will be applied in this work.

- Pass specific options for the optimizer internally by means of a control argument.

- Initial values

The algorithms for minimization functions require initial values of the parameters. There are several options to include initial values by means of ini.cov.pars argument:

- A unique initial value is used if a vector is provided in the argument ini.cov.pars. The elements are initial values for σ^2 (partial sill) and ϕ (range parameter), respectively. This vector is concatenated with the value of the argument nugget (nugget effect) if the argument fix.nugget is equal to false and the argument kappa if the argument fix.kappa is true.

- Specification of multiple initial values is also possible. If this is the case, the function searches for the one which minimizes the loss function and uses this as the initial value for the minimization algorithm. Multiple initial values are specified by providing a matrix in the argument ini.cov.pars and/or, vectors in the arguments nugget and kappa (if included in the estimation). If ini.cov.pars is a matrix, the first column has values of σ^2 and the second has values of ϕ.

- Alternatively the argument ini.cov.pars can take an object of the class eyefit or variomodel. This allows the usage of an output of the functions eyefit, variofit or likfit as initial value.

However, it must be remarked that:

- If the minimization function is nls, only the values of ϕ and κ (if this is included in the estimation) are used. Values for the remaining are not needed by the algorithm.

- If the selected model is the linear one, only the value of σ^2 is used. Values for the remaining are not needed by the algorithm.

- If the selected model is the pure nugget one, no initial values are needed since no minimization function is used in fact.

- Weights

The different options for the argument weights are used to define the loss function to be minimized. There are three available options for weights:

- npairs: indicates that the weights are given by the number of pairs in each bin.

- cressie: weights as those detailed for the weighted least squares estimator that was introduced at the beginning of this subsubsection. This is the option used in this work.

- equal: equal values for the weights; this corresponds to the ordinary least squares variogram fitting.

1.3.2.3. Ordinary kriging

Kriging is the most popular technique for spatial prediction. The krige.conv function of geoR package allows to perform different types of kriging: simple kriging, ordinary kriging, external trend kriging and universal kriging.

Ordinary krigging assumes that the spatial process Z can be decomposed in the sum of a constant mean μ and an intrinsic stationary process ε with a known semiovariogram γ, that is,

$$Z(s) = \mu + \varepsilon(s).$$

Then best linear unbiased prediction under square-error loss is known as ordinary kriging. The linear predictor in an arbitrary point s_0 is $p(Z;\ s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$, where we will set $\sum_{i=1}^{n} \lambda_i = 1$ to make it unbiased. Ordinary kriging consists of the selection of the best predictor that minimizes the mean squared error of the predictions

$$\min_{p} \mathbb{E}[(Z(s_0) - p(Z;\ s_0))^2]$$

Solving the previous minimization problem, the best predictor p_{ok} is obtained, and can be expressed in terms of semiovariogram as follows

$$p_{ok}(Z;\ s_0) = \left(\gamma + \mathbf{1}\,\frac{(1 - \mathbf{1}'\Gamma^{-1}\gamma)}{\mathbf{1}'\Gamma^{-1}\mathbf{1}}\right)' \Gamma^{-1} Z,$$

where $\gamma = (\gamma(s_1 - s_0), \cdots, \gamma(s_n - s_0))'$, Γ is the $n\ x\ n$ matrix which element (i, j) is $\gamma(s_i - s_j)$, and $\mathbf{1} = (1, \cdots, 1)'$.

The variance of the prediction can be expressed as

$$\sigma_{ok}^2(s_0) = \gamma'\Gamma^{-1}\gamma - \frac{(\mathbf{1}'\Gamma^{-1}\gamma - 1)^2}{(\mathbf{1}'\Gamma^{-1}\mathbf{1})}.$$

1.3.3. Key points

The basic ideas behind the geostatistical technique applied in this work, are briefly recapitulated in the following points (Schabenberguer & Gotway, 2005):

- The key in spatial data analysis is the **autocorrelation** of observations in space. The effects of autocorrelation in statistical inference is based on the idea of that n correlated observations do not provide the same amount of information than uncorrelated observations: high values are surrounded by high values, and low values are surrounded by low values.

- **Interpolation** is the estimation of a variable at an unmeasured location from observed values at surrounding locations. All interpolation algorithms (splines, radial basis functions, triangulation) estimate the value at a given location as a weighted sum of data with value at surrounding locations. Almost all assign weights according to functions that give a **decreasing weight with increasing separation distance**.

- **Kriging** can be understood as a two-step process: first, the spatial covariance structure of the sampled points is determined by fitting a variogram; and second, weights derived from this covariance structure are used to interpolate values for unsampled points. This assumes the process can be decomposed in the sum of a constant average value plus an intrinsic stationary process with known **semivariogram**.

2. OBJECTIVES

1. To test geostatistical interpolation techniques (kriging) in the study of a complex disease that displays heterogeneous spatial distribution (multiple sclerosis) as a tool to integrate information from different sources and at a different spatial resolution, explore the interactions between known risk factors and reproduce its spatial patterns.

    1.1. Characterize the geographic gradient of multiple sclerosis in Europe.

    1.2. Analyze the relative importance of the vitamin D-related factors included in the study.

2. To test geostatistical interpolation techniques (kriging) as an alternative to conventional genetic imputation, with an emphasis in the performance for low frequency variants.

3. APPLICATION 1: MULTIPLE SCLEROSIS IN EUROPE

3.1. MULTIPLE SCLEROSIS (MS) AS A CASE STUDY

3.1.1. Clinical aspects of MS

3.1.1.1. MS symptoms

Multiple sclerosis is a chronic complex disease of the Central Nervous System (CNS) that arises through a cell-mediated autoimmune process. What primarily triggers this autoimmune attack remains unclear yet leads to a gradual impairment of nerve signaling and, eventually, to an irreversible axonal damage (disability). It constitutes the first cause of non-traumatic neurological disability among young adults of Western Caucasian ancestry.

MS patients show discrete areas of neuroinflammation, demyelination, axonal degeneration, and gliosis -plaques- distributed across the CNS. Symptoms occur depending on the affected CNS region, being the most frequent first signs of the disease the following (MSIF and WHO, 2013):

- Sensory: numbness, paresthesia, Lhermitte's sign[4], pins and needles sensation, Uhthoff's phenomenon[5].
- Motor: useless hand syndrome[6], stiffness, spasticity, tremor, dystonia, ataxia, nystagmus, dysarthria.
- Visual: decreased vision or temporary blindness, scotoma, diplopia, blurred vision.

[4]Lhermitte's sign (barber chair phenomenon): electrical sensation that runs down the back and into the limbs (either up or down the spine) when bending the neck forward.
[5]Uhthoff's phenomenon: temporary worsening of neurologic symptoms caused by the increase of body temperature (caused by fever, hot weather, exercise, sauna, or hot tubs).
[6]Useless hand of Oppenheim: sudden numbness and awkwardness of one arm in which the sense of posture is most seriously affected.

Other early manifestations such as pain and fatigue, cognitive alteration -affecting information processing, decision making, verbal fluency, working memory, attention and concentration-, loss of balance and coordination, gastrointestinal issues, and urinary or sexual disorders are equally common, but may not be first directly associated with MS, yielding a delay in diagnosis.

3.1.1.2. MS courses

To provide a clear view on the clinical management and unequivocal defined checkpoints for clinical studies, MS has been categorized into several subtypes based on how disability accumulates over time. The classification has been refined over the years (Lublin and Reingold 1996; Lublin et al. 2014; Thompson et al. 2017), currently describing six general MS courses (as illustrated in Figure 4): the clinically isolated syndrome (CIS), the radiologically isolated syndrome (RIS), relapsing-remitting MS (RRMS), secondary progressive MS (SPMS), primary progressive MS (PPMS) and progressive-relapsing MS (PRMS).

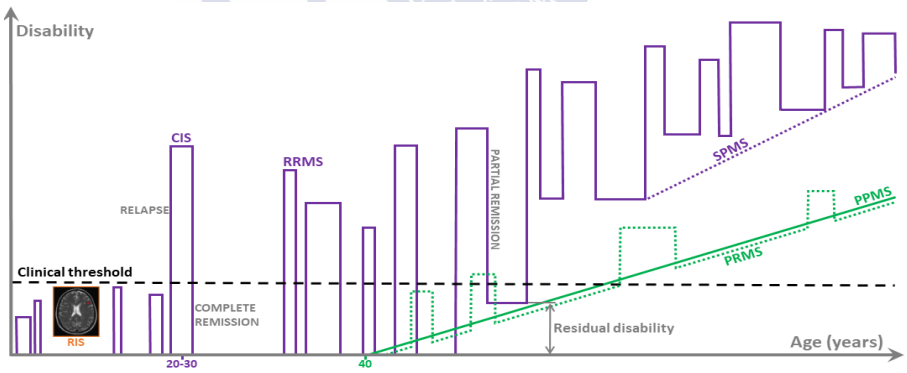

Figure 4. MS courses based on accumulation of disability over time. RIS: radiologically isolated syndrome. CIS: clinically isolated syndrome. RRMS: relapsing-remitting MS. SPMS: secondary progressive MS. PPMS: primary progressive MS. PRMS: progressive-relapsing MS.
MRI image is licensed under the Creative Commons Attribution-Share Alike 4.0 International license: author James Heilman, MD.

The clinically isolated syndrome (CIS) is the first episode of neurological deficit, defined as a 'monophasic clinical episode with patient-reported symptoms and objective findings, reflecting a focal or multifocal inflammatory demyelinating event in the CNS, developing acutely or subacutely, with a duration of at least 24 hours, with or without recovery, and in the absence of fever or infection' (Thompson et al., 2018). Subjects that experiment a CIS are more likely to develop MS than general population, and the second and successive episodes will be referred to as relapses, exacerbations, bouts, flare-ups, or attacks. Noteworthy, as illustrated in Figure 4, most MS affected individuals actually experiment episodes before CIS (a prodromal phase), that go unnoticed given that the clinical threshold is not surpassed (Dendrou et al., 2015).

Because of the incorporation of magnetic resonance imaging (MRI) techniques into regular clinical practice, the radiologically isolated syndrome (RIS) was defined. RIS consists in an incidental MRI finding that is suggestive of inflammatory demyelination but occurs in the absence of clinical signs or symptoms. Subjects with RIS are as well at a higher risk of developing MS.

In the relapsing-remitting course (RRMS, the most frequent presentation of MS) the disease occurs disruptively, with clearly defined attacks followed by a complete or partial remission after which patients remain at a stable period with no or almost any progression until the next relapse arises (either a new symptom or a relatively fast worsening of the current situation).

In nearly 80% of RRMS cases, and usually 15-20 years after the onset, the recovery from the relapses starts to be gradually more incomplete, originating a permanent accumulation of disability. MS evolves thus into a secondary progressive course (SPMS, secondary progressive MS).

Finally, primary progressive MS (PPMS) is characterized by a steady accumulation of disability over time since the very onset of the disease. In a low proportion of progressive cases, there are

superimposed exacerbations and remissions, resulting in the progressive-relapsing course (PRMS, progressive-relapsing MS).

Despite this general classification of disease subtypes, it should be noted that MS clinical course is heterogeneous in terms of activity and progression. The disease is considered active when a relapse arises and/or there is evidence of recent MRI lesions. Likewise, progression occurs when there is objective evidence of disease worsening over time, with or without relapses, or new MRI activity. Combining both aspects, MS evolution over time could vary from a mild illness (full recovery from sporadic episodes of neurologic symptoms) to a rapidly evolving and incapacitating one (frequent attacks with partial o no recovery).

In view of these circumstances and given the high variability of symptoms, the current trend considers MS a continuum/spectrum disorder, although the stratification into categories or endophenotypes (the six already outlined or future others) is necessary to carry out an efficient clinical management and research.

3.1.1.3. Diagnosis

First described in the XIX century, MS cases have been successively characterized to define and clarify the pathognomonic signs and symptoms that would facilitate their diagnostic and clinical management.

Early based on the detailed observation of patient's manifestations (e.g. Charcot's neurologic triad[7]) and partially driven by pharmaceutical clinical trials, the efforts on this matter have mainly focused on achieve consensus about diagnostic criteria, valid systems of assessment and effective therapeutic approaches, gradually incorporating the technological advances and research findings of their time.

[7]Charcot's neurologic triad: combination of nystagmus, intention tremor, and scanning or staccato speech.

An example of the latter, the development of the Experimental Autoimmune Encephalomyelitis (EAE) animal model (Yager, 1949) deserves a special mention, as it helped to understand the autoimmune origin of MS, apart from being the first step in testing the pharmacological activity of potential treatments.

It was relatively recently when the first agreed conclusions about the basic aspects of MS, such as the definition of attack, remission, and evidence of a lesion, were captured (Poser et al., 1983). This first effort was continued in the next decade, when Lublin and Reingold (Fred D; Lublin & Reingold, 1996) conducted an international survey to unify MS-related terminology among clinicians, laying the foundations of the course classification that, with some adjustments, remains in effect nowadays.

Five years later, the International Panel on MS Diagnosis stated the importance of dissemination of MS lesions in time (DIT) and space (DIS), always supported by objective clinical evidence of attacks or progression: MS lesions must develop over time and in different CNS locations (Mcdonald et al., 2001). The from then on called 'McDonald Criteria' emphasized the importance of a context-dependent differential diagnosis in order to rule out other non-MS disorders and formally incorporated magnetic resonance imaging (MRI), analysis of cerebrospinal fluid (CSF) and evoked potentials (EP) as essential tools in the diagnostic scheme.

MS lesions are typically be found in four CNS areas: periventricular (Dawson fingers[8]), juxtacortical and infratentorial regions, and in the spinal cord. MRI scanning not only detects demyelinated lesions in the CNS, but also allows their dating. Used with gadolinium contrast, MRI permits to distinguish between recent active lesions (current or up to 3 months inflammation regions) and aged ones (irreversible damaged regions). To note, lesions detected on MRI do not

[8] Dawson fingers: periventricular demyelinating plaques distributed along the axis of medullary veins, perpendicular to the body of the lateral ventricles and/or callosal junction. This is thought to reflect perivenular inflammation.

always correlate with patient symptoms, and more lesions do not necessarily imply a more severe MS-related disability.

In addition, the presence of CSF-specific oligoclonal bands (OCB) or elevated IgG index indicate abnormal immune response within the CNS. The analysis is considered positive when at least two oligoclonal bands not present in serum are found in CSF, there is an elevated IgG index, or both events occur (Brownlee et al., 2017).

Finally, EP tests measure the electrical activity of the brain in response to stimulation of a specific sensory nerve pathway. They can detect the slowing of electrical conduction caused by demyelination along these pathways even when the change is too subtle to be noticed by the own patient or to show up on a neurologic examination.

Both CSF analysis and EP tests help in the diagnostic process but are non-specific of MS. Until today, MRI is the preferred method to establish a diagnosis of MS and to monitor the course of the disease over time.

A later review of the 'McDonald Criteria' (C. Polman et al., 2005) clarified some aspects related to MRI, CSF analysis and acknowledged the need to achieve full validation of the Criteria in other populations (different to the Northern European) or in unusual forms of the disease.

The revision of 2011 (C. H. Polman et al., 2011) recognized the vital importance of translating patient's self-reported symptoms into objective measurements to achieve a proper diagnosis. It detailed the application of the Criteria in pediatric and Asian and Latin American populations, as well as the special considerations to take when there is risk of misdiagnosis with neuromyelitis optica spectrum disorders (NMOSDs).

The accuracy of MS phenotypes was the main topic of the 2013 and 2018 revisions (Fred D Lublin et al., 2014; Thompson et al., 2018). Table 1 summarizes the most recent consensus on MS diagnostic criteria.

	Number of lesions with objective clinical evidence	Additional data needed for a diagnosis of multiple sclerosis
≥2 clinical attacks	≥2	None
	1 (as well as clear-cut historical evidence of a previous attack involving a lesion in a distinct anatomical location)	None
	1	Dissemination in space demonstrated by an additional clinical attack implicating a different CNS site or by MRI
1 clinical attack	≥2	Dissemination in time demonstrated by an additional clinical attack or by MRI OR demonstration of CSF-specific oligoclonal bands
	1	Dissemination in space demonstrated by an additional clinical attack implicating a different CNS site or by MRI AND Dissemination in time demonstrated by an additional clinical attack or by MRI OR demonstration of CSF-specific oligoclonal bands.

Table 1. MS diagnostic criteria, 2017 revision (Thompson et al., 2018).

Nonetheless, the process of refining the MS diagnostic continues, as there is an increasing number of cases that do not meet full Criteria. For instance, in 'solitary MS' patients develop a progressive form of MS, but they possess a single cerebral, cervicomedullary junction or spinal cord lesion (with no clinical or radiological evidence of new lesion formation over time). For their part, late-onset and pediatric MS have an unusual age of debut and the increase in disability and the gravity of symptoms often differ from the observed in the already described MS subtypes. It is thought that the mechanisms of pathogenicity are somehow different, given the special characteristics of the affected subjects: elder people with age-related comorbidities and children with an immature immune system and fewer opportunities of prolonged exposure to risk factors.

In line with this, a current concern is that numerous medical conditions have overlapping symptomatology with MS, increasing the

risk of misdiagnosis. To mention a few: brain tumors, infections (Lyme disease, syphilis, viral infections), inflammatory disorders (vasculitis, Behçet's disease, SLE[9], Sjögren's syndrome), genetic diseases (hereditary myelopathies, leukodystrophies, CADASIL[10]), copper or vitamin B12 deficiency, structural CNS damage (herniated disc, cervical spondylosis, Chiari's malformation), and other demyelinating disorders (NMOSD, ADEM[11]). This is why the research in the field has been prioritizing the discovery of a MS blood or CSF biomarker that either confirms the condition (Baranzini, Srinivasan, et al., 2010; Haghighi et al., 2013; Al-Temaimi et al., 2017) or correlates with the triggering of relapses (Milosevic et al., 2015), the activity of the disease or the response to current pharmacological treatments (Harris & Sadiq, 2014; Signoriello et al., 2016; Hewes et al., 2017; Tatomir et al., 2017).

3.1.1.4. MS treatment

Current MS therapeutic management takes place principally at two levels: the treatment of the acute symptoms during attacks with corticosteroids (dexamethasone, methylprednisolone, ACTH[12]), and the long-term treatment with disease-modifying therapies (DMT)[13], that reduce relapse rate and slow down disease progression (and the disability coupled to it). Of them, first-choice treatments are beta-interferon (1a, 1b) and glatiramer acetate (GA), followed by humanized monoclonal antibodies (natalizumab, ocrelizumab and alemtuzumab) and fingolimod.

Other EMA[14]-approved pharmacological agents for MS, such as dimethyl fumarate, mitoxantrone, teriflunomide, or cladribine are therapeutic alternatives used in severe refractory cases or in advanced

[9] SLE: systemic lupus erythematosus
[10]CADASIL: Cerebral Autosomal-Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy.
[11]ADEM: Acute Disseminated Encephalomyelitis.
[12]ACTH: Adrenocorticotropic hormone.
[13]DMT: also called immunomodulatory therapies (IMT), disease modifying agents (DMA) or disease modifying drugs (DMD).
[14]EMA: European Medicines Agency.

stages of the disease (due to their unspecific activity and sometimes serious secondary effects, i.e. cardiotoxicity of mitoxantrone).

Overall, the therapeutic advances of recent years allowed to achieve a much better prognosis for treated patients in relapsing-remitting MS, reducing both severity and frequency of new exacerbations (in some cases, up to 60% reduction of relapse rate; but there is a marked variability in the response to treatments). By contrast, the fewer available treatment options for the progressive forms reached comparatively unsatisfactory results so far, probably due to still unidentified pathogenic mechanisms.

3.1.2. Epidemiology of MS

MS affects roughly 2.8 million people worldwide, with a global median prevalence of 33 per 100,000 (MSIF and WHO, 2013). This value has increased with respect to the previous report -30 per 100,000- (MSIF and WHO, 2008), possibly due to the harmonization of diagnostic criteria and the improvement of health systems around the world. Even so, many countries do not have robust systems to monitor MS yet, although efforts are currently being made to collect and unify available information in registries (Flachenecker et al., 2014).

Global prevalence has risen in the last few decades, although varies considerably between world regions, as presented in Figure 5. In general, MS prevalence is higher in North America and Europe (values of 140 and 108 per 100,000 respectively), and lower in Sub-Saharan Africa and East Asia (2.1 and 2.2 per 100,000 correspondingly).

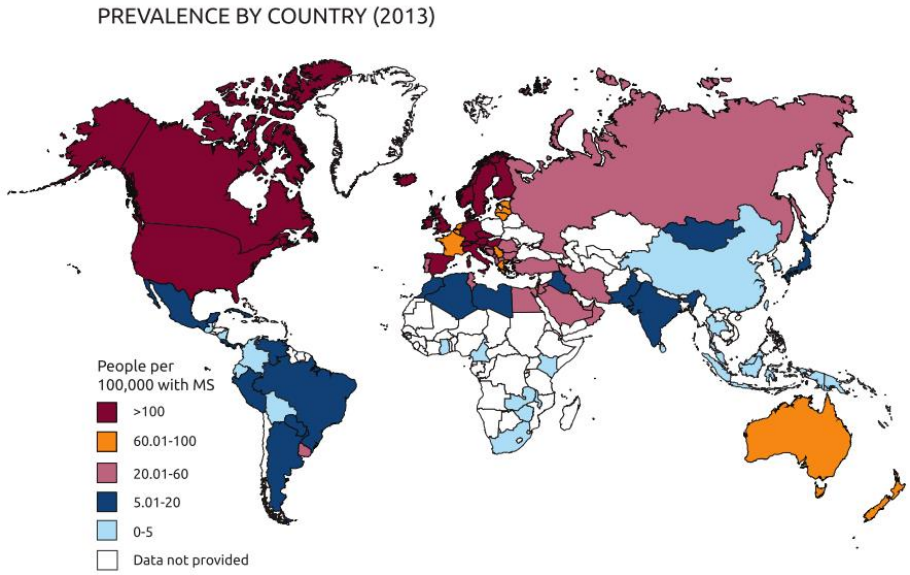

Figure 5. Worldwide prevalence of MS by country, with permission of The MS International Federation (MSIF and WHO, 2013).

Most of the affected individuals present a relapsing-remitting course (85%), including patients that eventually will develop a secondary-progressive form of the disease. Only 10% of subjects are diagnosed with a primary progressive course, and even less (5%) have progressive-relapsing MS from onset.

Over the years, several epidemiological factors have been related to MS risk: gender, familial relatedness, Northern European ancestry, genetic variants, and environmental risk factors. They will be briefly described next:

3.1.2.1. Gender

Except for the primary progressive course that equally affects both sexes and has a late age of onset, MS is considered a young female disorder (average age of onset 20-40 years). In Europe, female-to-male ratio varies from 1.22 in Cyprus to 3.60 in Portugal (3.1 in Spain) (MSIF and WHO, 2013).

Meta-analyses suggest that the incidence of MS has increased over time and provide some evidence that this has primarily resulted from an increase in the incidence of MS among women (Kingwell et al., 2013). This affirmation has been validated in Swedish (Boström & Landtblom, 2015) and Japanese populations (Houzen et al., 2018), although the concrete role of gender in the physiopathology of the disease is still under study (Golden & Voskuhl, 2017; Avila et al., 2018).

3.1.2.2. Familial relatedness

Studies found some MS cases showed familial aggregation: the risk among siblings varies by the degree of relatedness and the number of affected individuals within the family (Isobe et al., 2013). In general, it is thought that first-degree relatives could have 7-25 times greater risk of developing MS than the general population (Willer et al., 2003; O'Gorman et al., 2012; Westerlind et al., 2014).

However, the MS inheritance model behind this remains unclear. In monozygotic twins there is a concordance rate of only 20-30%, whereas in heterozygotic twins this rate does not surpass 5%. Baranzini et al. incorporated mRNA transcriptome and epigenetics in twin's studies without uncovering any reproducible difference between co-twins (Baranzini, Mudge, et al., 2010).

3.1.2.3. Northern European ancestry

It is known that countries with majority of Northern European ancestry population tend to have higher MS prevalence (Pugliatti et al., 2006; Kingwell et al., 2013; Flachenecker et al., 2014).

Likewise, studies have shown lower prevalence of MS in African Americans: African American men had approximately 40% lower MS risk than individuals with Northern European ancestry (Munger et al., 2006; Sawcer et al., 2011; Wallin et al., 2012). Also, African Americans tend to have more aggressive forms of the disease compared to Caucasian patients, and present differences in clinical manifestations:

more frequent opticospinal MS and transverse myelitis development (Cree et al., 2004).

3.1.2.4. Genetic variants

To date, roughly 200 independent genetic associations have been found in MS: 161 SNP-trait associations in the NHGRI-EBI GWAS catalog (Buniello et al., 2019). Still, genetic associations discovered so far are not strong enough to state a causal relationship, as current knowledge links only 30% MS risk to 'purely' genetic factors (Dendrou et al., 2015).

To note, MS-related genetic variants are shared by other autoimmune conditions, such as type 1 diabetes mellitus, SLE, rheumatoid arthritis, or Crohn disease; although their effect can be opposite depending on the disease (Baranzini & Oksenberg, 2017).

In addition, the MS genetic associations identified in the last decade are characterized by a progressive decrease in size effect, lately finding variants of low to moderate effect (Sawcer et al., 2011; Baranzini & Oksenberg, 2017).

Among the hits, the strongest signal is in the class II Human Leukocyte Antigen (HLA) (Ramagopalan et al., 2009). The HLA locus encodes the Major Histocompatibility Complex (MHC) set of membrane glycoproteins, participating in antigen processing and other aspects of host defense. HLA-DRB1 is expressed in antigen-presenting cells and is relevant in the regulation (both activation and inactivation) of CD4 T cells. Several HLA haplotypes have been associated with MS (with either a risk or a protective role), as well as their interactions (Beecham et al., 2013; Patsopoulos et al., 2013; Moutsianas et al., 2015).

Notably, the HLA-DRB1*15:01 allele, in the region 6p21, increases MS risk in a dose-effect manner: heterozygosity confers an OR of 2.91 (95%CI 2.42-3.51), and homozygosity of 5.42 (95%CI 4.12-7.16) in Northern European populations (Ramagopalan et al.,

2010). Unlike other DRB1 alleles, the structure of 15:01-binding groove has been shown to present both myelin and EBV[15] peptides to T cells (Kumar et al., 2013), fact that could potentially drive autoimmune reactivity (Tschochner et al., 2016).

Individuals having the risk allele for HLA-DRB1*15:01 with history of infectious mononucleosis are at increased risk of MS (OR 7.32, 95%CI 4.92-10.90). Although they are independent risk factors, there is a certain additive interaction that points towards a shared pathological mechanism between this HLA allele and EBV (Disanto et al., 2013).

Other HLA alleles with smaller effect have also been linked to MS: HLA class II DRB1*03:01 (OR 1.26), DRB1*13:03 (OR 2.40) and HLA class I A*02:01 (OR 0.73) (M. Marrosu et al., 1997; Sawcer et al., 2011).

MHC regions implication in the disease seems clear, even though their ultimate global effect is still under study due to its complex genetic architecture, a very likely epistasis between haplotypes (Beecham et al., 2013; Patsopoulos et al., 2013; Moutsianas et al., 2015) and their interaction with environmental MS risk factors, such as EBV, smoking (A. Hedström et al., 2017) and vitamin D (see below).

3.1.2.5. Environmental risk factors

Environmental MS risk factors have been progressively incorporated into ongoing prospective studies. Apart from the difficulty of getting standardization in measurements of exposure, it seems that MS risk might be influenced for both the duration of exposure and the timing in subjects' development: during gestation and neonatal period, childhood, adolescence, or early adulthood. Some of the most investigated environmental factors are the following:

[15] EBV: Epstein-Barr Virus, causal agent of the infectious mononucleosis.

- Infectious agents

The idea of a close relationship between MS and viral infections (Kurtzke, 2013) is still under study. It is thought that the underlying mechanism by which virus could trigger autoimmunity in MS is molecular mimicry, although this is still not well established (Geginat et al., 2017).

Of all the infectious agents, EBV is the most studied to date. EBV is a ubiquitous human herpesvirus that has ability to infect, activate, and latently persist in B lymphocytes for the lifetime of the infected subject; and it is the causal agent of infectious mononucleosis (IM).

Having high titers of antibodies anti-EBV has been linked to a higher risk of developing MS, and nearly the 100% of patients with MS are seropositive for EBV compared with the 90% of healthy people (Ebers, 2008). As aforementioned, HLA-DRB1*15:01 subjects with past EBV infection are at a higher risk of developing MS (Disanto et al., 2013).

Other infectious agents currently under research in MS are the Human Endogenous Retroviruses (HERV) (Christensen, 2017; Morandi et al., 2017; Tao et al., 2017). HERVs are remnants of ancestral retroviral infections throughout evolution. It has been found that HERV RNA and antigen levels are higher in MS patients (Mameli et al., 2012). For that reason, they could be potential biomarkers of MS, and are being studied in clinical trials (Dolei et al., 2019).

- Smoking

Smoking habit is a MS risk factor with a dose-response relationship and a OR 1.54 (95%CI 1.22-1.87) for current smoker vs past or never smoker status (Degelman & Herman, 2017).

Recently, Hedström et al. observed an interaction between smoking and HLA-DRB1*15:01, hypothesizing that smoke-induced lung-irritation may trigger autoaggressive T cells in the lungs or post-translationally modify peptides that are cross-reactive with CNS antigens, promoting a CNS-directed autoimmunity that results in MS (A. Hedström et al., 2017).

- Obesity

Several studies have detected association between MS and body mass index (BMI) in young adults (A. K. Hedström et al., 2012; Munger et al., 2013; Mokry et al., 2016; M. A. Gianfrancesco et al., 2017). Obese females in their 20's have 2-fold increased risk of MS compared normal-weighted ones (M. Gianfrancesco et al., 2014).

- Microbiome, diet

Microbiota (microorganisms colonizing human body; mostly non-pathogenic bacteria and fungi) is extremely important to maintain homeostasis. The microbial diversity and activity vary in a dynamic fashion, depending on factors such as diet, health status or pharmacological treatment. That is why one of the current trends in GxE interaction research is the study of microbiome (the combined genetic material of microbiota).

Gut microbiome is a known inflammation regulator, highly influenced by dietary habits (Riccio & Rossano, 2017). In MS, it has been found that affected individuals present qualitatively and quantitatively different microbiome composition (dysbiosis) with respect to healthy controls (Wang & Kasper, 2014; Chen et al., 2016; Mirza & Mao-Draayer, 2016). In particular, it is thought that during the age window of adolescence and young adulthood, dysbiosis could trigger the development of pathogenic, self-antigen–specific adaptive immune responses, playing a pathological role in the initiation and progression of MS (Yadav et al., 2017).

Research has linked a diet rich in polyunsaturated fatty acids (PUFA), found in aliments such as tree nuts -almonds, pecans, and pistachios-, dairy products, eggs, meat, shellfish, salmon, tuna or sardines, with lower MS risk (Bjørnevik et al., 2017). Two recent reviews of dietary intervention trials with PUFA in MS patients concluded that there is no clear evidence of less severe disease activity, amelioration of symptoms, lower relapse rate or overall clinical status (Mische & Mowry, 2018; Farinotti et al., 2020).

- Latitude

MS prevalence has a heterogeneous distribution worldwide (as previously illustrated in Figure 5): equatorial regions have lower prevalence than areas near the poles. This suggests a geographic classification of world regions by risk and latitude (Pugliatti et al., 2006; Beretich & Beretich, 2009) in high, medium or low risk areas (MSIF and WHO, 2013).

Interestingly, migration studies have found that migration to a higher risk country before age 15 makes subjects acquire higher MS risk, but this does not happen when the migration takes place later in life (Compston, 1997; Hammond et al., 2000).

At a smaller scale, some epidemiological studies have collected proof of a latitudinal gradient in prevalence or incidence of MS in Latin America (Risco et al., 2011), New Zealand (Alla & Mason, 2014) and France (Vukusic et al., 2007; Fromont et al., 2010).

Latitude has been significantly associated with both early age of MS onset (Tao et al., 2016) and MS prevalence (Simpson et al., 2011). With some reserves in the latter in the last years (Koch-Henriksen & Sorensen, 2011; Grant & Mascitelli, 2012), the latitudinal gradient of MS prevalence has been recently re-evaluated in a meta-analysis including age and sex standardized, HLA-DRB1-adjusted values of prevalence from 94 studies, and confirmed (Simpson et al., 2019).

- Sunlight exposure

Areas with yearly deficient sunlight access (latitudes above 40º) correspond with high MS prevalence areas. Studies found that higher sun exposure during childhood and early adolescence is associated with a reduced risk of MS (Van Der Mei et al., 2003; Tremlett et al., 2018).

Although sun exposure may have its own anti-inflammatory effect (Hart & Gorman, 2013) and thus reduce MS risk (Breuer et al., 2014; Langer-Gould et al., 2018), experimental and epidemiological data suggest that vitamin D is the predominant mediator of the sunlight effect.

- Vitamin D

Vitamin D (VD) is a fat-soluble secosteroid that promotes the intestinal absorption and metabolism of calcium and phosphorus (relevant in bone homeostasis) and has multiple other implications as a modulator of gene expression in numerous cell-types, including myeloid cells (Booth et al., 2016).

VD can be obtained either by food or supplement intake or by endogenous production. As Figure 6 illustrates, sunlight exposure (UV-B sunlight in the range of 290-315 nm) transforms 7-dehydrocholesterol into previtamin D3 in the skin, that spontaneously isomerizes into vitamin D3 (cholecalciferol). Cholecalciferol is then transported by the vitamin D binding protein (VDBP) into the liver, where undergoes a first hydroxylation by CYP27A1. Bind to the Alpha-globulin (AG), calcidiol travels to the kidney where it turns into the biologically active form after another hydroxylation by CYP27B1. VDBP will finally transport calcitriol to target cells, where VD acts as transcription regulator via nuclear receptors (VDR).

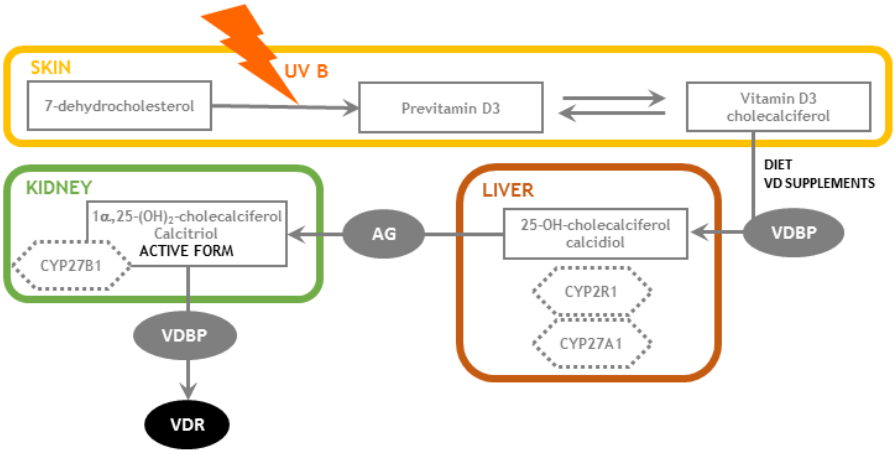

Figure 6. Metabolism of vitamin D.

Based on bone metabolism requirements, the recommended daily VD dose is 75-80 nm/L or 2000UI (Pierrot-Deseilligny & Souberbielle, 2010). In early MS studies, deficit in serum VD levels was found among

patients (Munger et al., 2006). This directed the research of VD out of its initially known functions, testing its role in MS physiopathology, and yielding to the confirmation of VD implication in the immune component of the disease (Hart & Gorman, 2013; Pierrot-Deseilligny & Souberbielle, 2013; Wöbke et al., 2014).

Moreover, VD supplementation has been tested with therapeutic intention, findings suggesting that could decrease MRI lesions (Ascherio et al., 2014) and reduce the risk of relapses (Simpson et al., 2010), but with inconclusive results at least until now. It has been discussed that the reason behind this could be that MS affected individuals are less responsive to VD supplementation compared to healthy controls (Bhargava et al., 2016).

Yet, VD remains undoubtedly as an attractive linker between genetics and environment in MS: several environmental factors participate in VD metabolism in a compatible manner with the spatial distribution of the disease (latitude, sun exposure, diet), and an immunomodulatory role of VD has been already confirmed: the VDRE in the promoter region of HLA-DRB1*15:01 (Correale et al., 2009; Kragt et al., 2009; Ramagopalan et al., 2009; Simon et al., 2011). The complete implications of this are not fully known but could reasonably involve a differential immunological milieu in MS patients that could interplay with other known risk factors (infectious agents, smoking). All without considering other possible VD functions that remain undiscovered.

3.1.3. Rationale

For the first application of geostatistical techniques to the study of CD, we prioritized the risk factors associated with MS described in the preceding sections that met the following criteria:

- To be related to the hypothesis of VD as linker between genetic and environmental factors.
- To display a well-established geographic trend.

- To be relevant in European populations.
- To be easily available, either from literature or public databases.

Figure 7 summarizes the genetic and environmental MS risk factors finally considered (colored), that will be the following:

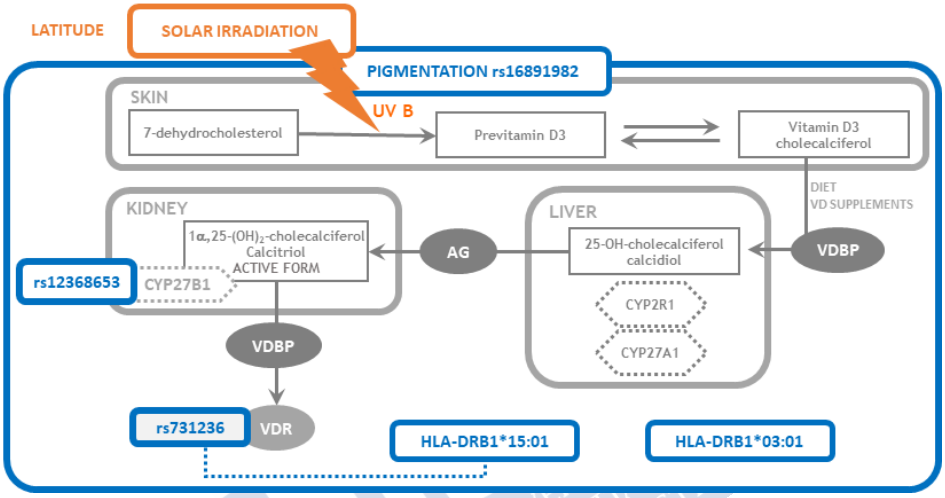

Figure 7. MS risk factors considered for this application: environmental (latitude, solar irradiation) and genetic (pigmentation rs16891982, HLA-DRB1*15:01, HLA-DRB1*03:01, CYP27B1 rs12368653, VDR rs731236).

- *Immunological component (HLA alleles)*

HLA-DRB1*15:01 is the main genetic MS risk factor in Caucasians and has a VDRE in its promoter. HLA-DRB1*03:01 is the second most relevant allele associated with MS risk after adjusting for HLA-DRB1*15:01 (Sawcer et al., 2011; Mokry et al., 2016), especially in Italian regions (M. Marrosu et al., 1997; M. G. Marrosu et al., 2001).

- *Solar irradiation*

This environmental factor is necessary for the endogenous production of VD. Plus, it is conditioned by latitude, as areas near the Equator have higher solar irradiation.

- *Skin pigmentation rs16891982 (allele G)*

The SNP rs16891982 is an ancestry informative marker (AIM) frequently used as a proxy of the gene SLC45A2, which determines skin pigmentation. The inclusion of this marker in our model was justified for two main reasons: first, the European ancestral allele (G) is associated with light skin and its frequency increases with latitude (skin pigmentation shows a latitudinal gradient), and second, lighter skin has been proven to be more efficient synthesizing VD (Jablonski & Chaplin, 2000, 2012; Yuen & Jablonski, 2010).

- *VDR rs731236 (allele T)*

VDR belongs to the steroid-thyroid-retinoid acid receptor superfamily, that binds to DNA to regulate transcription. It is the last player needed for VD to act as a transcription modulator. T allele has been previously associated with lower risk of MS in HLA-DRB1*15:01 positive individuals (Agliardi et al., 2011).

- *CYP27B1 rs12368653 (allele G)*

This gene encodes the hydroxylase that catalyzes the production of the biological active form of VD. It is located in a reported MS susceptibility region of chromosome 12 and has been confirmed as a candidate gene in the autoimmune mechanisms of the disease (Sundqvist et al., 2010). To include its possible influence in MS distribution, we used the SNP rs12368653 as a proxy of CYP27B1. The ancestral allele in Europe (G) is associated with a decreased MS risk (ANZ, 2009).

3.2. MATERIALS AND METHODS

3.2.1. Data

For reproducibility purposes, all datasets used for this application are included in Appendix 9.1 (A-F).

3.2.1.1. Prevalence data

MS prevalence (number of individuals affected per one hundred thousand people) from 117 European locations was collected from literature (Kingwell et al., 2013). When MS prevalence data for the same location was available in two different time periods, the most recent value was selected, leaving a final number of prevalence data of 109. When prevalence and age-standardized prevalence were available, the age-standardized value was preferred. Geographic coordinates (latitude and longitude) were manually assigned to each location.

3.2.1.2. HLA-DRB1*15:01 and HLA-DRB1*03:01 data

HLA data (at 92 locations for HLA-DRB1*15:01 and 90 for HLA-DRB1*03:01) was downloaded from the Allele Frequency Net Database (AFND), a publicly available online repository for immune gene frequencies in worldwide populations, including human leukocyte antigens, major histocompatibility complex class I chain-related genes, killer-cell immunoglobulin-like receptors, and cytokine gene polymorphisms (González-Galarza et al., 2015; Ghattaoraya et al., 2016).

3.2.1.3. Solar irradiation data

Long-term average global irradiation on horizontal plane ($Wh/m^2/day$) for each MS prevalence location (N=109) was downloaded from the Photovoltaic Geographic Information System (PVGIS), a publicly available database of solar irradiation and temperature historical series (Huld et al., 2012).

3.2.1.4. Pigmentation data - rs16891982 (G)

G allele frequency in 30 geographic locations was downloaded from the Melgene Database (Chatzinasiou et al., 2011; Athanasiadis et al., 2014; Antonopoulou et al., 2015) and ALFRED, The Allele Frequency Database (H. Rajeevan et al., 2003; Haseena Rajeevan et al., 2005, 2012). The first one is a publicly available field synopsis of genetic association studies in cutaneous melanoma; the second gathers allele frequencies for DNA sequence polymorphisms in anthropologically defined human populations (Europe, in our case).

3.2.1.5. VDR data - rs731236 (T)

Frequencies of the allele T for the SNP rs731236 in 49 European locations were collected from literature (see details Appendix 9.1. E).

3.2.1.6. CYP27B1 data - rs12368653 (G)

Frequencies of the allele G at 18 locations were downloaded from the HGDP, the Human Genome Diversity Project (Cavalli-Sforza, 2005). The objective of the HGDP initiative is to record the genetic profiles of indigenous populations to understand human migration and evolution.

3.2.2. Analysis

Even though we will be referring to MS prevalence as outcome and the variables included as predictors, it is worth remembering that the purpose for the regression model in this section is not inference, but a way of integrating information and explore interactions between the considered MS risk factors.

All the statistical analysis were performed in R (R Foundation for Statistical Computing, 2018).

3.2.2.1. Input data overview

To visualize the spatial distribution of the variables included in this analysis, prevalence data and the predictors (as.geodata objects) were plotted. This generates a four-panel graphic that allows to easily visualize geo-referenced data, and notice possible overall spatial trends (latitudinal, longitudinal): in the upper-left panel, a plot representing the geographic distribution of the input data, assigning different colors and shapes by quartiles; in the upper-right and bottom-left panels, scatterplots of values of data vs latitude and longitude, respectively; and in the bottom-right panel, a histogram and density plot of the data. Spearman correlations with latitude and longitude were also performed to verify any visually detected geographic trends.

To have a general overview of the setup for this application, input data of MS prevalence and risk factors were summarized in table, including the sample size of each variable, the median (interquartile range), the expected effect in MS based on current knowledge, any geographic trend detected, and minimum and maximum values, as well as the geographic locations (country) associated to them (see Section 3.3.1.8).

3.2.2.2. Kriging

An initial assessment of the spatial properties (isotropy, stationarity and spatial dependence) of the predictors was performed following published methods (Diblasi & Bowman, 2001; Adrian W. Bowman & Crujeiras, 2013), with the sm.variogram function in sm R package (A. W. Bowman & Azzalini, 2018). Shapiro-Wilk test was used to assess normality. Alpha level of significance was set at 0.05.

As kriging can be affected by this, transformation of MS prevalence and rescaling of predictors were also tested, but as the results did not change noticeably, original values were kept for an easier interpretation.

To make use of all data available of the predictor variables, the jittering of the geographic coordinates was carried out when two values refer to the same geographic location (jitterDupCoords function).

Empirical variograms were calculated and used in ordinary kriging to obtain estimates of the predictors at the prevalence locations. In the case of CYP27B1 SNP rs12368653 (G), an adjustment to a robust empirical variogram with lineal tendency was performed first.

3.2.2.3. Correlation

Spearman correlation test of the kriged values and actual MS prevalence data was carried out. A correlogram was plotted with corrplot R package (Wright, 2018).

3.2.2.4. Model

In view that the Krige-and-regress (K-R) approach (Madsen et al., 2008) yielded a complex and hard to interpret model (see Section 3.3.5); an additional step of Principal Component Analysis (PCA) before regression was applied (K-PCA-R).

PCA is a dimensional reduction method based on the generation of orthogonal linear combinations of variables (called principal components or PCs) that account for as much variability in the data as possible. Although relying in the scale of variables, this method is handy to combine information of highly correlated data (as is our case). The number of PCs to include as terms in the model was decided viewing the cumulative proportion of variance they accounted for (at least 85 % of the variability in the data).

The model including all possible two by two interactions between the PCs underwent a backward stepwise selection of terms based on Akaike Information Criterion (AIC). Briefly put, the AIC is a measure of the relative quality of a model: it balances the goodness-of-fit with the model complexity (number of terms). The summary of the selected model was displayed and fitted values calculated.

3.2.2.5. Visual representations

The selected PCs were plotted by pairs, labelling the original MS prevalence point data locations by country/region, using factoextra R package (Kassambara & Mundt, 2020).

In order to facilitate their visualization, point locations were colored in k=5 groups as a result of hierarchical clustering based on their distances, defined with functions hclust and cutree from stats R package (R Foundation for Statistical Computing, 2018). Several numbers of clusters were tested, but it was finally decided to use k=5, as the groups created are easily followed through the plots and define actual European geographical regions (see Section 3.3.4). Grouping by prevalence ranges was also tested, but no clear differentiation was found.

Contour maps of the real and predicted values of MS prevalence (see Section 3.3.6) were generated in R, using libraries geoR (Ribeiro Jr & Diggle, 2018), maps (Becker et al., 2018) and ggplot2 (Wickham, 2016).

3.3. RESULTS

3.3.1. Input data overview

3.3.1.1. MS prevalence in Europe

There is a latitudinal gradient in MS prevalence ($rho_{Spearman}$=0.54), that partially decreases Eastwards ($rho_{Spearman}$ with longitude=-0.22). As already showed in Figure 5 and clear in the bottom right density plot, the European area of study is a high prevalence region (most values above 50 cases/100000).

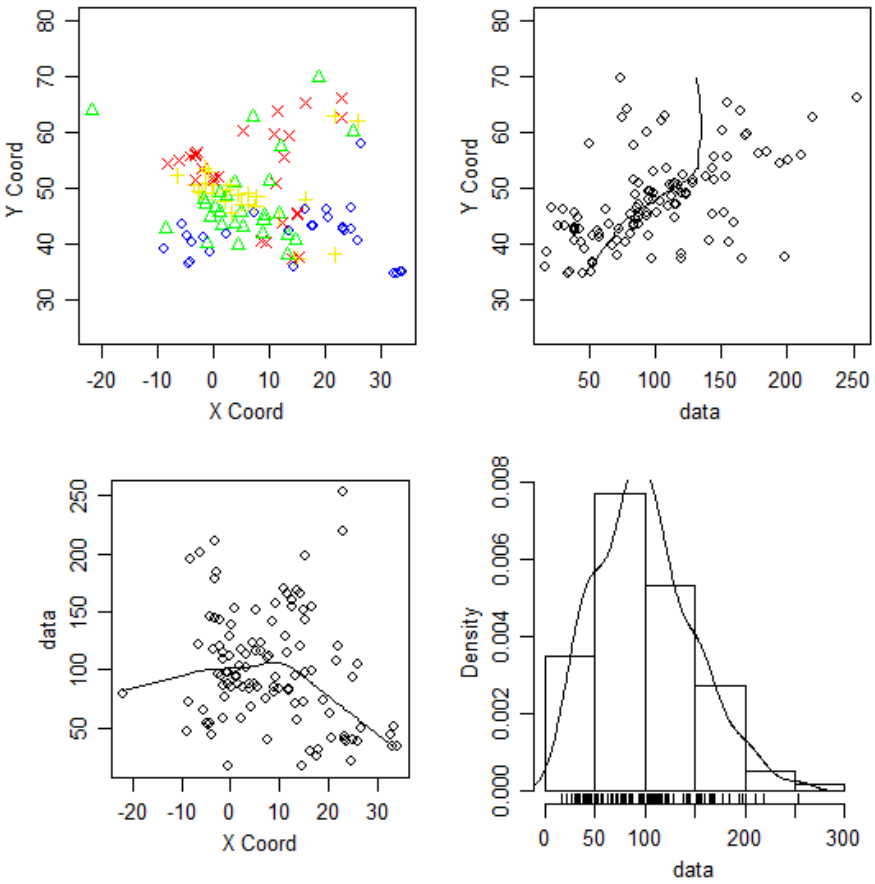

Figure 8. Plots of MS prevalence (N=109). Y Coord=latitude. X Coord=longitude.

3.3.1.2. HLA-DRB1*15:01

To note the uneven distribution of input data, with a concentration of points in the British Islands-Ireland and in Iberian Peninsula. There is a positive correlation between HLA-DRB1*15:01 and both longitude and latitude ($rho_{Spearman}$ =0.29, $rho_{Spearman}$ =0.59 respectively) meaning that, in general, Northeastern regions in our area of study have higher values of HLA-DRB1*15:01 frequency. Most data are between in the ranges 0.06-0.09 and 0.11-0.15.

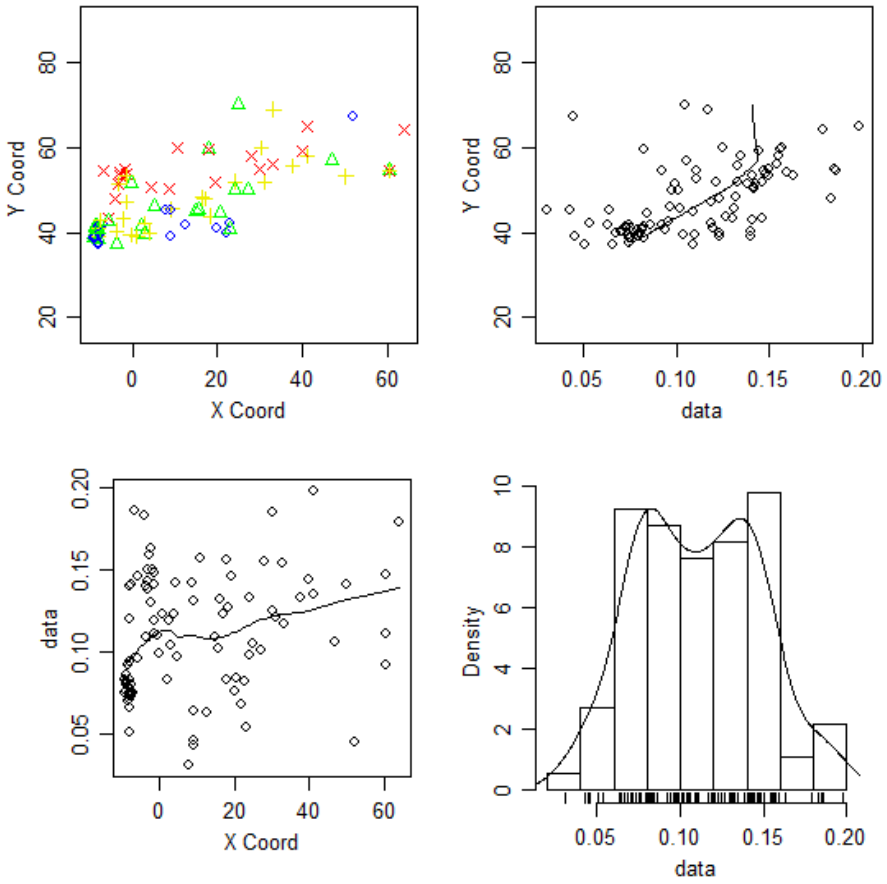

Figure 9. Plots of HLA-DRB1*15:01 data (N=92). Y Coord=latitude. X Coord=longitude.

3.3.1.3. HLA-DRB1-03:01

Again, there is a denser concentration of similar valued data in the British Islands and the Iberian Peninsula. This could interfere in the found negative correlations with longitude ($rho_{Spearman}$ =-0.55) and latitude ($rho_{Spearman}$ =-0.25), not previously reported: Southwestern regions in our area of study tend to have higher frequencies for HLA-DRB1*03. Frequencies for this allele concentrate roughly in the range 0.05-0.20, with an extreme value in Sardinia.

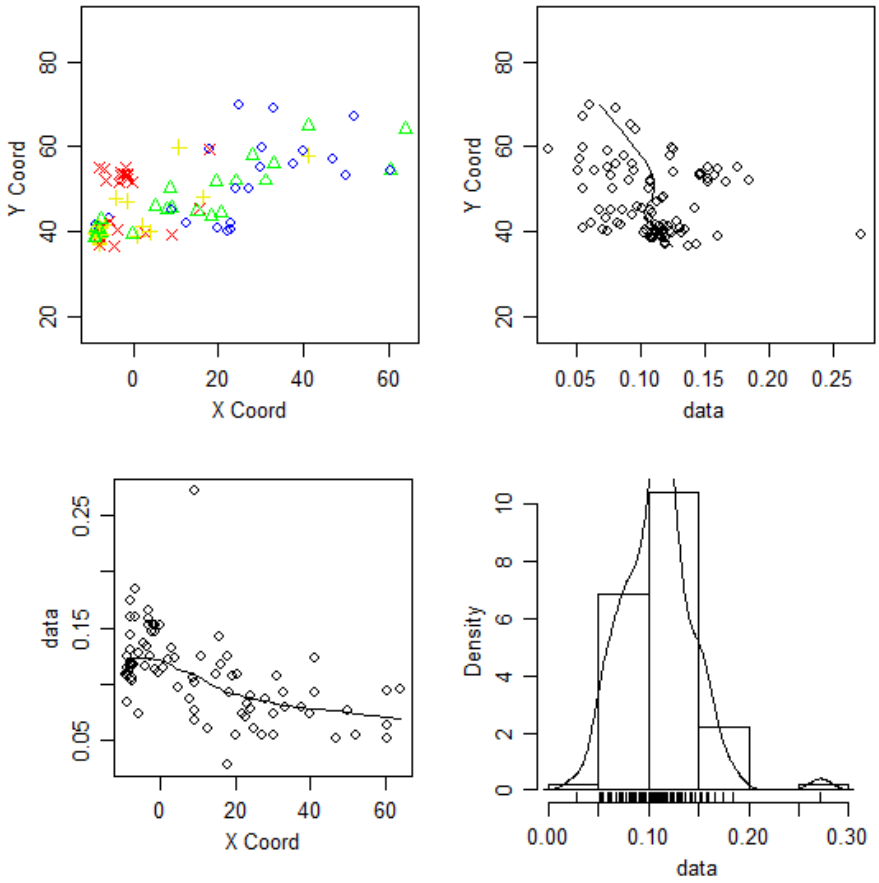

Figure 10. Plots of HLA-DRB1*03:01 data (N=90). Y Coord=latitude. X Coord=longitude.

3.3.1.4. Solar irradiation (Wh/m^2/day)

As expected, there is a strong latitudinal gradient in solar irradiation ($rho_{Spearman}$=-0.98). Yearly average data oscillate between 2000 Wh/m2/day in Northern areas and almost 5000 in the Mediterranean regions.

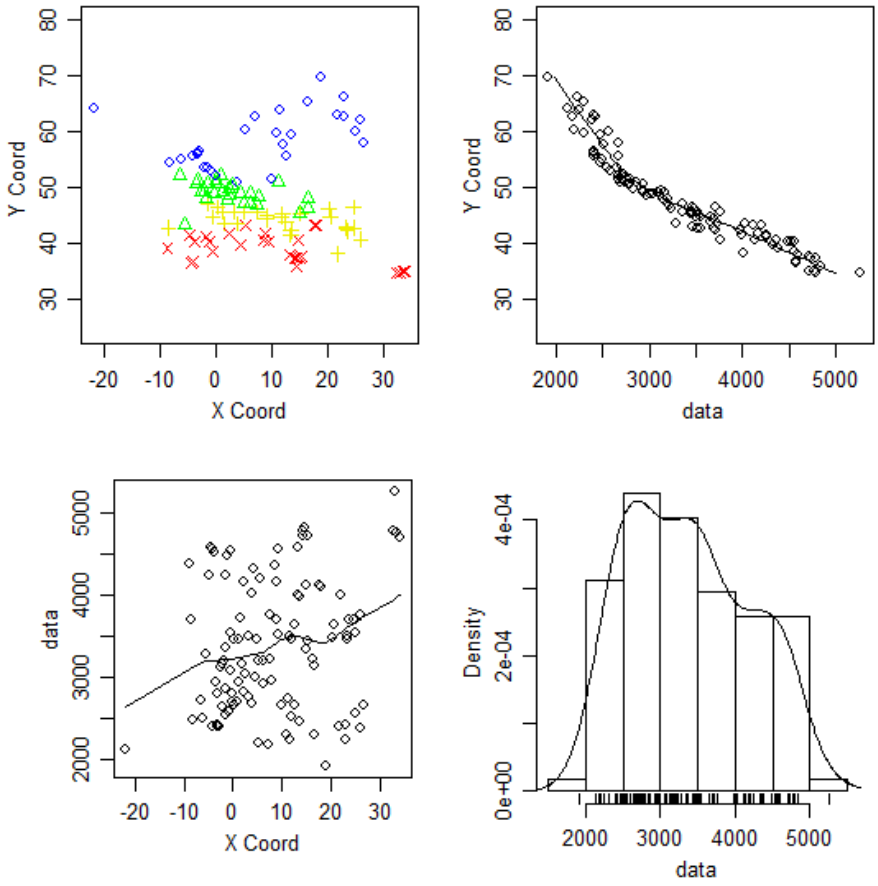

Figure 11. Plots of solar irradiation data (N=109). Y Coord=latitude. X Coord=longitude.

3.3.1.5. Pigmentation SNP: rs16891982 (G)

Even with the modest number of data, a latitudinal gradient is observed in the allele G frequency of rs16891982 (linked to light skin, $rho_{Spearman}$ =0.57), as previously established in the literature (Jablonski & Chaplin, 2000). Most data are in the range 0.75-0.99 of frequencies, except an Italian location (Sardinia) that has the lowest value.

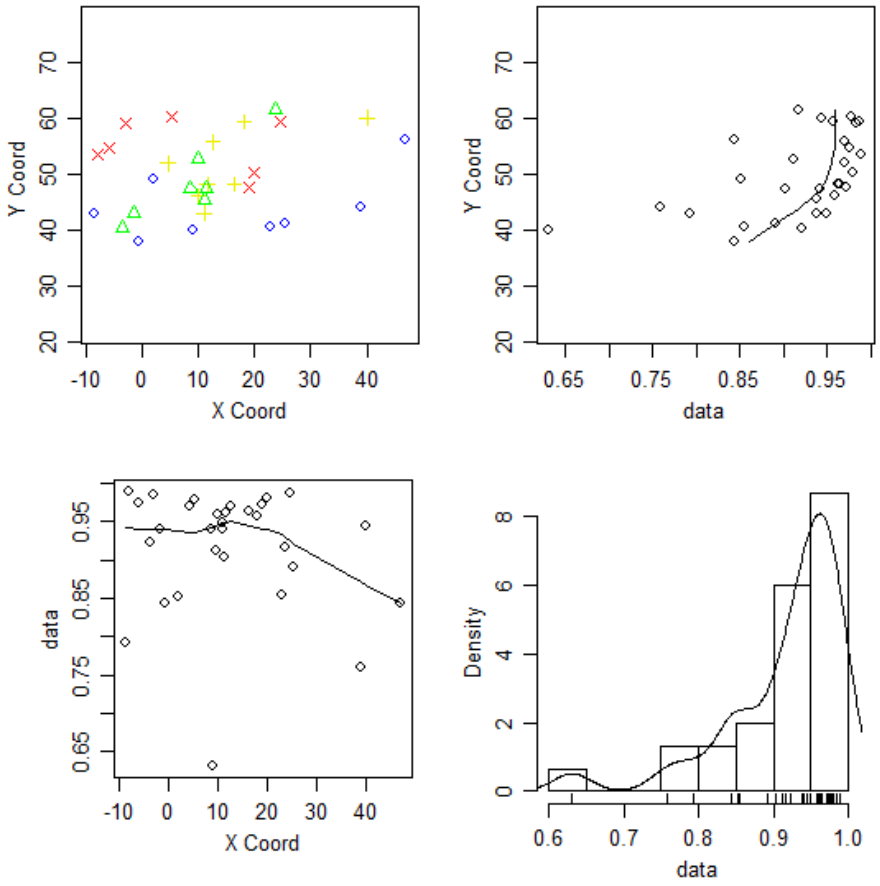

Figure 12. Plots of pigmentation data (N=30). Y Coord=latitude. X Coord=longitude.

3.3.1.6. VDR SNP: rs731236 (T)

No correlation of T allele frequency of rs7321236 with longitude or latitude was significant. Most frequencies are >0.55, except Crete, that has the minimum value.

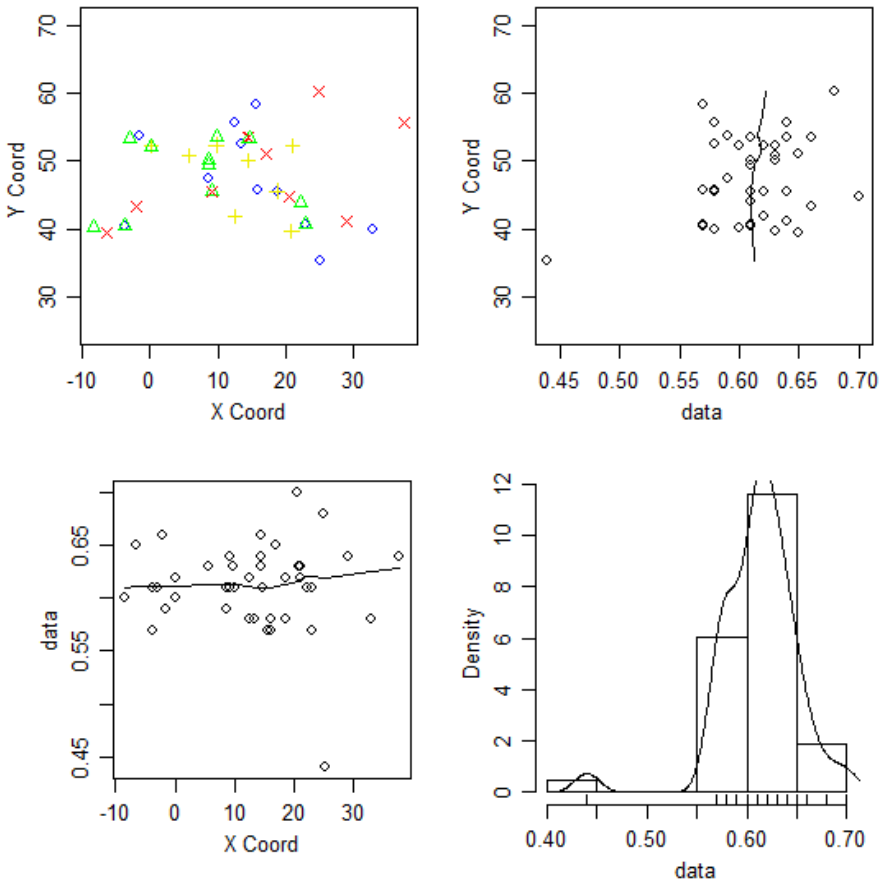

Figure 13. Plots of VDR data (N=42). Y Coord=latitude. X Coord=longitude.

3.3.1.7. CYP27B1 SNP: rs12368653 (G)

Higher latitudes have higher G allele frequencies for the CYP27B1 SNP rs12368653 ($rho_{Spearman}$ =0.54), with most values in the range 0.50-0.60. This would imply that the protective allele frequency is higher in Northern regions (the ones with high MS prevalence), where prevalence is high. This could be due to the limited data sample (N=18).

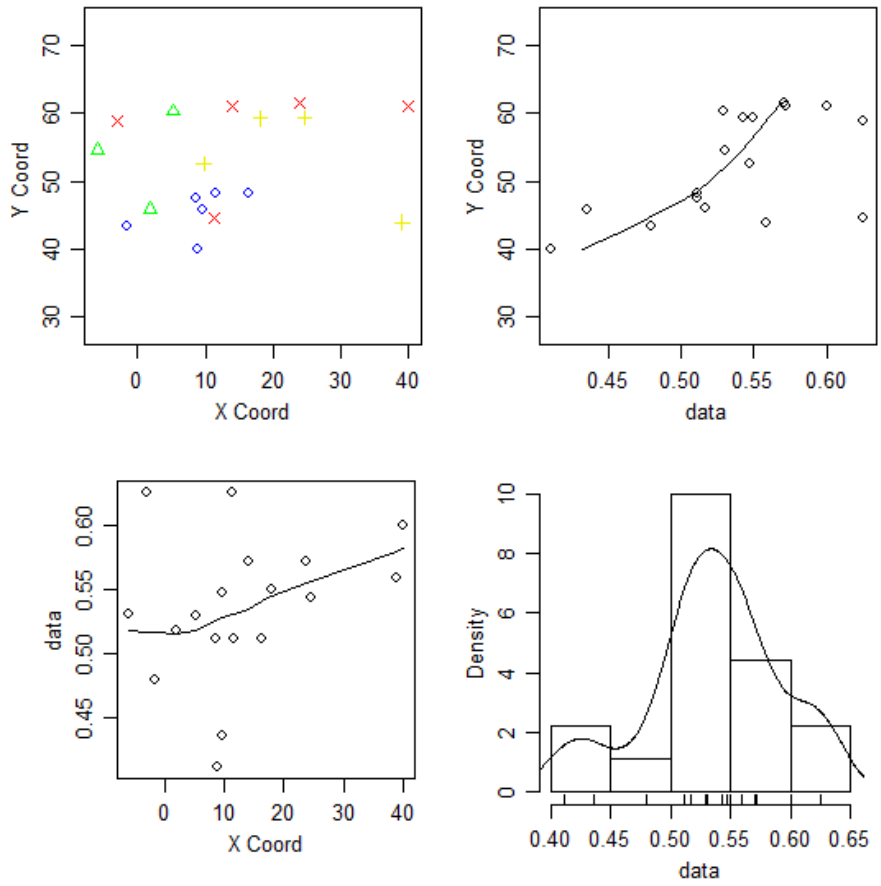

Figure 14. Plots of CYP27B1 data (N=18). Y Coord=latitude. X Coord=longitude.

3.3.1.8. Summary table

Table 2 summarizes the input data for this application:

Variable	N	Median (IQR)	Effect in MS	Geo trend	Minimum (location)	Maximum (location)
MS prevalence	109	94.7 (58.7)		+LAT -LONG	16.7 (Malta)	253 (Sweden)
HLA-DRB1*15:01	92	0.11 (0.06)	R	+LAT +LONG	0.03 (Italy)	0.20 (Russia)
HLA-DRB1*03:01	90	0.11 (0.04)	R	-LAT -LONG	0.03 (Finland)	0.27 (Italy)
Solar irradiation (Wh/m2/day)	109	3350 (1340)	-	-LAT	1920 (Norway)	5260 (Cyprus)
Pigmentation rs16891982 (G allele)	30	0.94 (0.08)	-	+LAT	0.63 (Italy)	0.99 (Ireland and Estonia)
VDR rs731236 (T allele)	43	0.61 (0.04)	P	-	0.44 (Crete)	0.70 (Serbia)
CYP27B1 rs12368653 (G allele)	18	0.54 (0.06)	P	+LAT	0.41 (Italy)	0.62 (Italy and Scotland)

Table 2. Description of the input data for the application 1.
N: sample size.
IQR: interquartile range.
Effect in MS: evidence in the literature; P=protective effect; R=risk effect
Geo trend: +LAT=positive correlation with latitude, +LONG=positive correlation with longitude, -LAT=negative correlation with latitude, -LONG=negative correlation with longitude.

3.3.2. Kriging

Results from spatial conditions assessment are summarized in Table 3. As previously described (see Section 1.3), kriging performs better when data are homogeneous (isotropic and intrinsically stationary), normal and spatial dependent. These assumptions are advisable but not strictly required, so the fact that some predictors do not fit all the properties does not invalidate our results. Moreover, some of these tests are conservative, so it will be difficult for out limited set of data to comply with the conditions. Thus, we carried out ordinary kriging to estimate the risk predictors at the locations where MS prevalence data are available (N=109).

Predictor variable	Isotropy p-value	Intrinsic stationarity p-value	Normality p-value	Spatial dependence p-value
HLA-DRB1*15:01	0.385	0.319	0.1766	0.063
HLA-DRB1*03:01	0.588	0.111	8.57e-4	0.066
Solar irradiation	0.133	0.131	1.94e-3	0.018
PIGM - rs16891982 (G)	0.29	0.026	4.71e-5	0.13
VDR - rs731236 (T)	0.66	0.002	1.48e-4	0.527
CYP27B1 - rs12368653 (G)	0.005	0.549	0.535	0.171

Table 3. Spatial condition assessment: isotropy (H_0: isotropic), intrinsic stationarity (H_0: stationary), normality test (Shapiro-Wilk, H_0: normality), spatial dependence (H_0: independent). Alpha level of significance = 0.05.

3.3.3. Correlations of the kriged data

As anticipated, MS prevalence is positively correlated with latitude and inversely correlated with solar irradiation (higher prevalence at Northern regions, where solar irradiation is low). Known MS risk factors HLA-DRB1*15:01 and HLA-DRB1*03:01 display a positive correlation with prevalence, agreeing with the current knowledge that links higher MS risk in carriers of those HLA alleles.

The cline of pigmentation is exhibited in our limited dataset, with a very strong negative correlation of pigmentation rs16891982 (G) with solar irradiation (meaning lighter skin in low irradiation areas).

Regarding the general geographic trends that were observed in the input data overview (see Section 3.3.1.8), there is a close match with the correlations now calculated, with only minor inconsistencies in very weak correlations with latitude (of HLA-DRB1*03:01 and VDR) and longitude (HLA-DRB1*15:01).

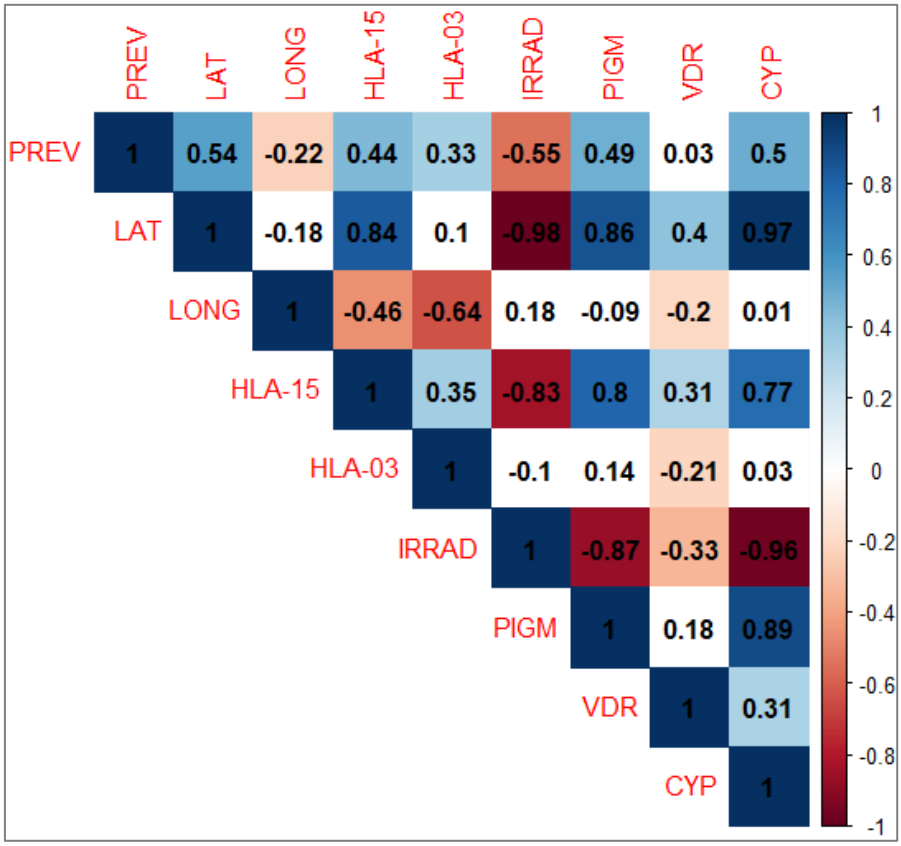

Figure 15. Correlogram of real MS prevalence data and the kriged predictor variables. Numbers indicate the Spearman correlation coefficient. Non-significant coefficients at alpha level 0.05 are colored blank. PREV=MS prevalence. LAT=latitude. LONG=longitude. HLA-15=HLA-DRB1*15:01. HLA-03=HLA-DRB1*03:01. IRRAD=solar irradiation. PIGM=rs16891982(G). VDR= rs731236(T). CYP= rs12368653(G).

To note, kriged CYP27B1 rs12368653 (G) values are extremely correlated with latitude and irradiation. Due to the low number of collected CYP27B1 rs12368653 (G) data (N=18) and the limited range of variation in the allele frequency, the standard approach of kriging resulted in constant kriged values, that is why we adapted the original ordinary kriging methodology. It seems that there are not enough data to confidently describe a spatial pattern (that differs from the latitudinal, that is already collected in other predictors). Similarly, latitude provides redundant information (mimicked by solar irradiation). For these reasons, CYP27B1 rs12368653 (G) and latitude will be excluded of further analysis.

3.3.4. Principal component analysis (PCA)

As showed in Table 4, principal component 1 (PC1) is mostly composed by HLA-DRB1*15:01, solar irradiation, and pigmentation. PC2 is based on longitude and HLA-DRB1*03:01 allele. Finally, PC3 is defined mostly by the frequency of VDR. The three first PCs explain >85% of variance in the predictors.

Principal component	LONG	HLA-15	HLA-03	IRRAD	PIGM	VDR	Eigenvalue	% variance
PC1	-0.27	0.54	0.13	-0.53	0.50	0.31	2.98	49.6 %
PC2	0.58	-0.10	-0.70	-0.17	0.22	0.31	1.62	79.6%
PC3	0.36	0.18	0.06	-0.11	0.37	-0.83	0.77	89.6%

Table 4. Eigenvectors of the first 3 PCs, eigenvalues and cumulative proportion of variance explained (% variance).

Biplots for the 3 PCs were created, adding color to facilitate the visualization (as previously described in Section 3.2.2.5).

- PC2 vs PC1, Figure 16:

The graphical representation of PC2 versus PC1 consistently gathers point locations by actual European geographical regions. Five non overlapping groups are easily identified: Scandinavian regions in blue, East (Italy) and West Mediterranean regions (in red and orange, respectively), a central European group (in purple) and a North-

Northwestern cluster (British Islands, in green). The distribution of clusters in the plot fairly corresponds to the real geographic locations, as PC1 is mostly composed by predictors with a marked latitudinal gradient, and PC2 itself is majority longitude and HLA-DRB1*03:01 with extreme values of frequency in Italy and Finland.

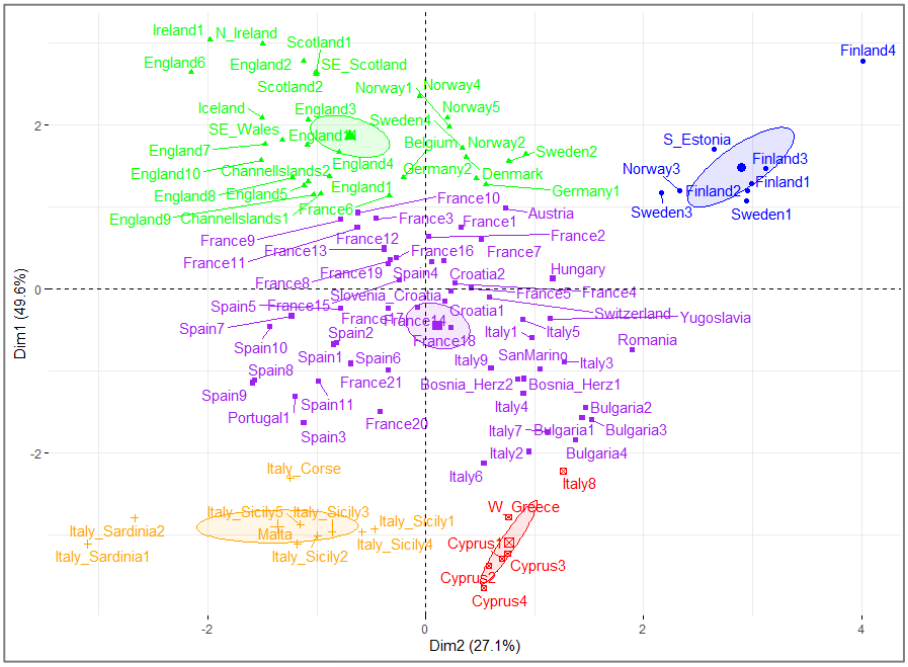

Figure 16. Representation of PC2 vs PC1. PC1 was put in the Y-axis to facilitate the interpretation of the plots.

- PC3 vs PC1, in Figure 17:

PC3 is composed mainly by the VDR rs731236 (T), followed by pigmentation and longitude. Again, PC1 represents mostly latitude, although the Northern regions are swapped when comparing with the previous plot.

- PC2 vs PC3, in Figure 18:

In this case, the three biggest clusters overlap; only Eastern Mediterranean and the Scandinavian regions seem to clearly differentiate from the rest.

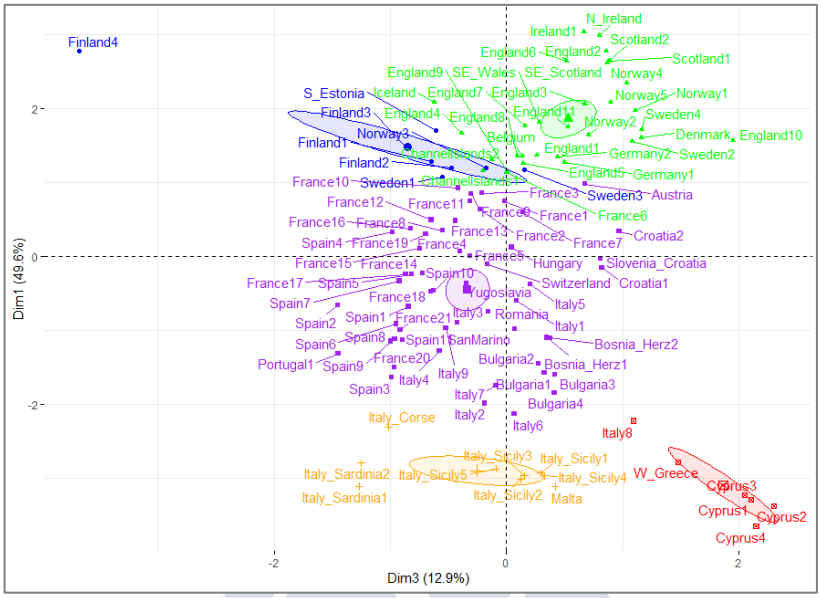

Figure 17. Representation of PC3 vs PC1. As PC1 is composed by terms with latitudinal gradient was put in the Y-axis, as it will facilitate the visualization.

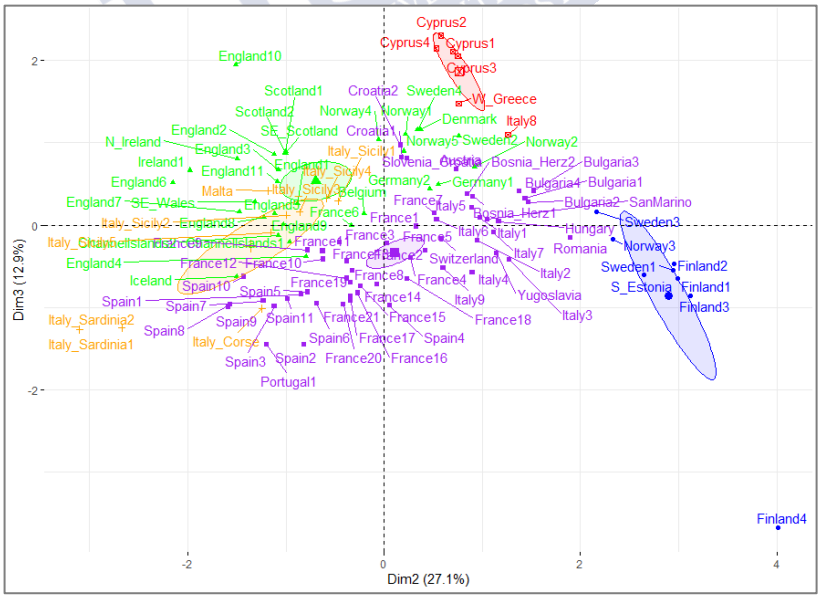

Figure 18. Representation of PC2 vs PC3.

3.3.5. Models

Table 5 illustrates the complex model resulting from K-R.

adjR2=0.4661				
Term	**estimate**	**std.error**	**statistic**	**p-value**
(Intercept)	-6.05E+04	2.63E+04	-2.297	0.024
LONG	4.13E+01	1.78E+01	2.316	0.023
IRRAD	1.73E+00	7.16E-01	2.415	0.018
HLA.15	-5.53E+04	2.64E+04	-2.095	0.039
HLA.03	-1.49E+03	9.35E+02	-1.588	0.116
PIGM	6.93E+04	2.94E+04	2.361	0.020
VDR	9.40E+04	4.26E+04	2.208	0.030
LONG:PIGM	-4.73E+01	1.97E+01	-2.401	0.018
IRRAD:HLA-03	4.65E-01	2.59E-01	1.798	0.075
IRRAD:PIGM	-8.09E-01	2.34E-01	-3.461	0.001
IRRAD:VDR	-1.76E+00	1.05E+00	-1.671	0.098
HLA-15:VDR	8.91E+04	4.31E+04	2.067	0.041
PIGM:VDR	-1.07E+05	4.75E+04	-2.255	0.026

Table 5. Model resulting from the K-R approach.

This first approach leads to a complex model will difficult interpretation. For this reason, K-PCA-R model was selected instead (Table 6) for further analysis and map representation.

adjR2=0.3251				
Term	**estimate**	**std.error**	**statistic**	**p-value**
(Intercept)	100	3.83	26.13	<2.0E-16
PC1	10.67	3.23	3.31	1.3E-03
PC2	-5.03	3.67	-1.37	1.7E-01
PC3	20.85	5.62	3.71	3.4E-04
PC1:PC2	6.06	2.00	3.03	3.1E-03
PC1:PC3	5.83	2.62	2.23	2.8E-02
PC2:PC3	-5.13	3.49	-1.47	1.4E-01

Table 6. Model resulting from the K-PCA-R approach.

The model K-PCA-R of the predictor risk factors would explain roughly one third of the variability of MS prevalence data. The composition of PCs (see previous section) is consistent with the individual predictors selected with the previous K-R approach. This means that independently of the approach applied, the same predictors are selected.

There is a strong correlation ($rho_{Spearman}$=0.64, p=1.065e^{-13}) between K-PCA-R fitted values and observed MS prevalence, as Figure 19 illustrates.

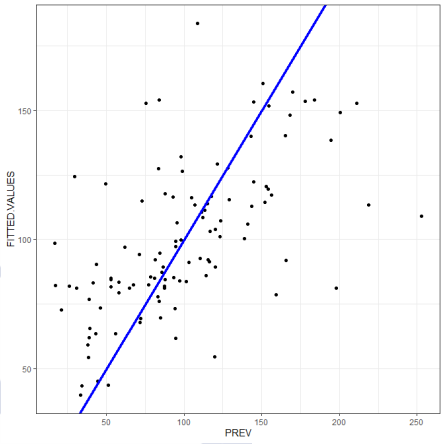

Figure 19. Observed MS prevalence vs predicted values by the model.

3.3.6. Contour maps

Contour maps of the real values of MS prevalence (Figure 20) and the predicted values using the K-PCA-R model (Figure 21) show an overall concordance, capturing the main latitudinal trend combined with a minor NW-SE gradient.

Both plots successfully identify high MS prevalence areas in the North (British Island, Nordic countries) and in the Sardinian region, although the K-PCA-R map results in a smoothed trend surface.

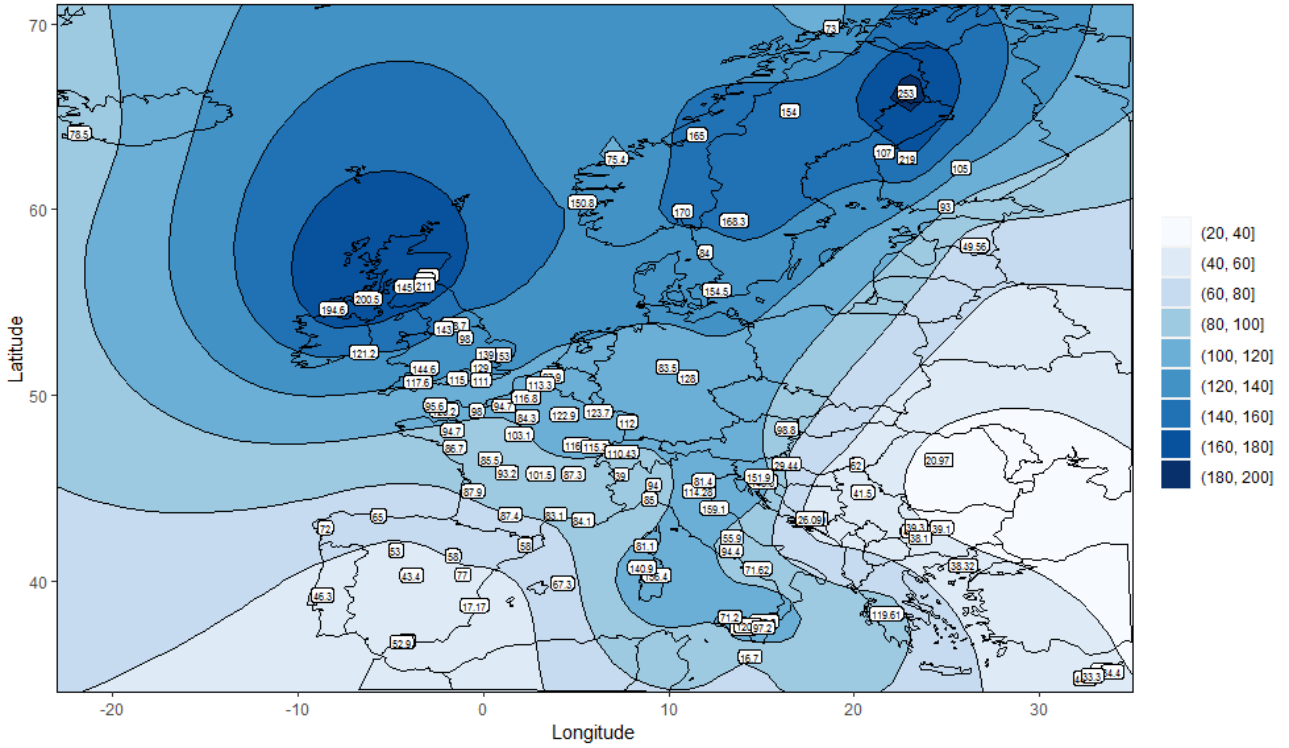

Figure 20. Map of real MS prevalence data included in the study. Labels with prevalence values superimposed.

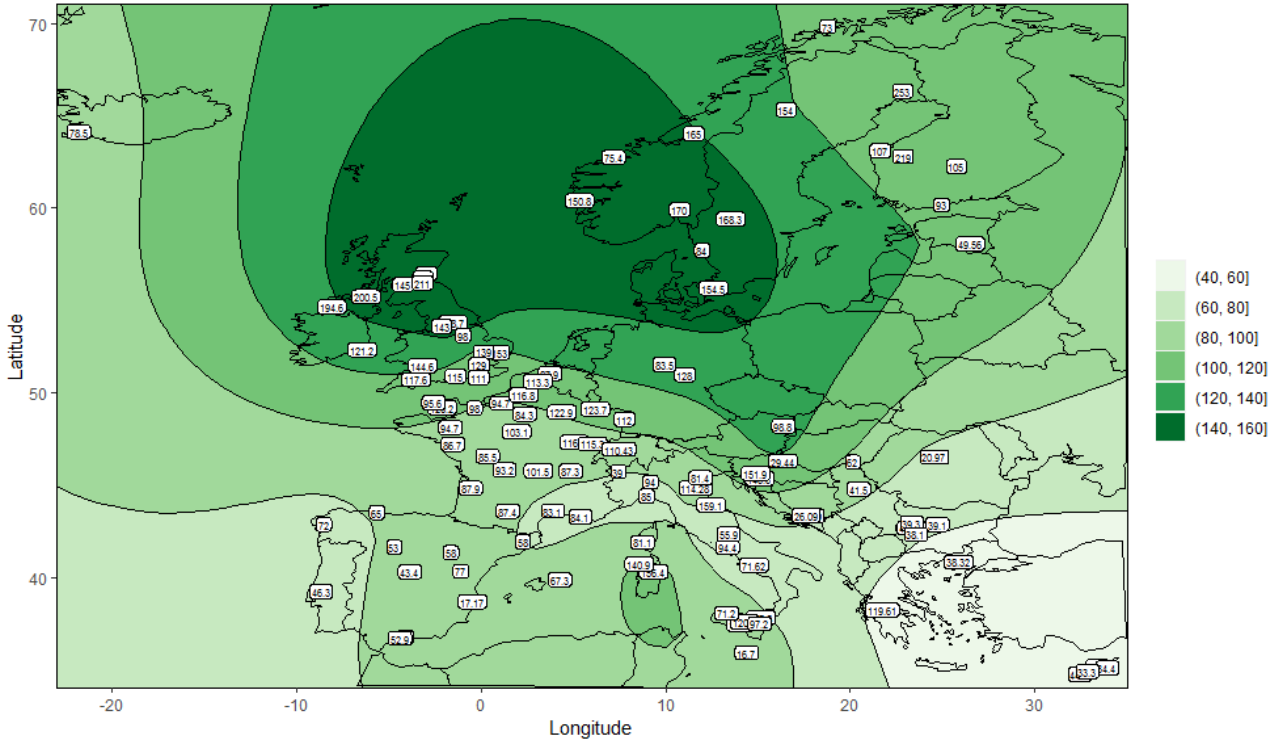

Figure 21. Map of predicted MS prevalence with the K-PCA-R model. Labels with real prevalence superimposed.

4. APPLICATION 2: KRIGING VS GENETIC IMPUTATION

4.1. GENETIC IMPUTATION

4.1.1. Overview

The use of genetic imputation tools is currently a regular step in the normal workflow of genetic association studies and metanalyses, due to the variable composition of markers in the arrays of commercial platforms. At present, there are several imputation programs available: SHAPEIT (Delaneau et al., 2012, 2014; Delaneau, Howie, et al., 2013; Delaneau, Zagury, et al., 2013)-IMPUTE2(Marchini et al., 2007; B. N. Howie et al., 2009; Marchini & Howie, 2010; B. Howie et al., 2011, 2012), MACH (Li et al., 2009, 2010)-minimac (B. Howie et al., 2012; Fuchsberger et al., 2015), and BEAGLE (Browning & Browning, 2007). SHAPEIT and MACH are phasing tools for IMPUTE2 and minimac, respectively.

Although each one has its own particularities, they are all based mechanistically in the use of Hidden Markov Models (HMM) to estimate haplotypes prior to imputation. Briefly put, HMM are probabilistic models of sequential data. In this case, the sequence is the sequence of typed SNPs[16], and the probabilities are calculated based on linkage disequilibrium (LD) and recombination rate of positions in the data and the reference haplotype.

Even though the theoretical approach is simple, the computational cost is high. It was not until recently (Michigan Imputation Server (Das et al., 2016)) that an high-performance computing (HPC) environment

[16] In this application, the terms SNP, variant and marker will be used indistinctly.

was required for carrying out genetic imputation. Figure 22 summarizes the very basic components needed for genetic imputation.

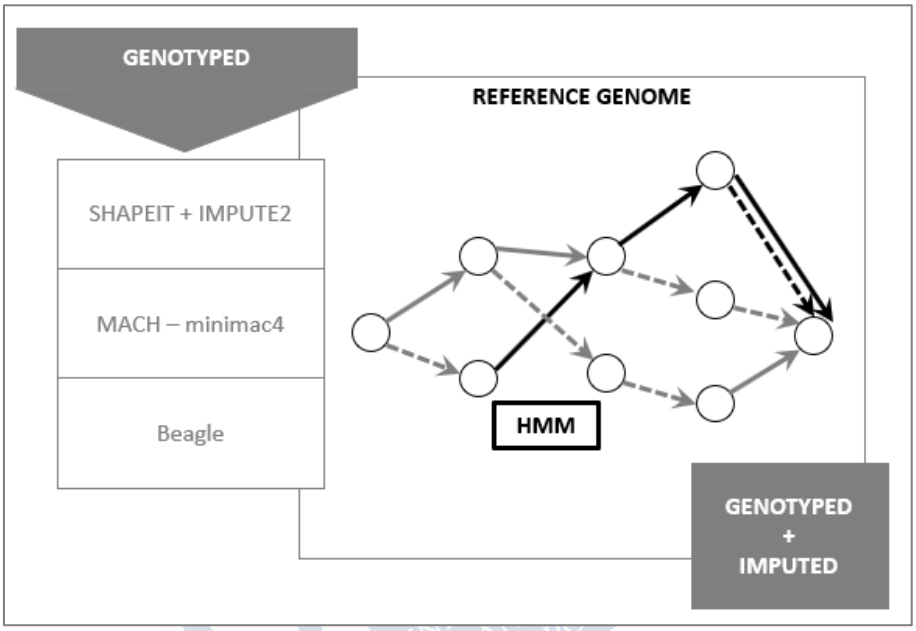

Figure 22. Diagram of genetic imputation methods. HMM=Hidden Markov Model.

The general imputation process will be described next:

First, genotyped data usually undergo a pre-imputation quality control (QC) step, depending on the characteristics of the array used for genotyping and the experimental design. The aim is to identify samples or markers that were not correctly genotyped (missing rate above a threshold or deviation from normal proportions of heterozygosity). It is essential that the genotyped and reference genome are in the same build/version of the reference genome. Otherwise, genotyped and reference population markers will not match.

After choosing the software and the reference population (for example 1KP), is time to ensure that input file formats are adequate. Depending on the software, additional files would be required, i.e.

recombination rate per chromosomal position. To be noted, only markers present in the reference genome will be imputed.

Usually the imputation takes place in two steps: the first one estimates the phase of haplotypes and the parameters of the model; then the second refines and predicts genotypes.

Finally, output files including genotyped and imputed SNPs, and some measure of quality of the imputation is generated. For instance, r2 of MACH is the ratio of the empirically observed variance of the allele dosage to the expected binomial variance at Hardy Weinberg Equilibrium (HWE); BEAGLE indicates the R2 between the best guess-genotype and the true genotype, and IMPUTE2 provides two measurements related to the relative statistical information about the marker frequency from the imputed data (info metric and certainty) (Marchini & Howie, 2010). Before using imputed data, a post-imputation QC step is advisable to remove markers with weak imputation results.

4.1.2. Limitations

Genetic imputation software was first designed to predict frequency of common SNPs with an additive/linear effect. The algorithms behind it are based on robust weights calculated by comparison of the present SNPs in both sample and reference haplotype, and only untyped positions that are present in the reference haplotype will be imputed in the sample.

Thus, the main conditioning factor is the **reference population**: this includes the number of markers to be imputed, the weights and haplotypes considered, the recombination rate, etc.

Depending on the software, the selection of the reference population can be user-defined or program-defined. For example, in our case, IMPUTE2 has a black box system in which it carries out contrasts

of each sample with the 1KGP reference data. It will choose the reference haplotype that best fits each single sample. That implies that the user does not know which reference population was used, not even if the same population group was used for all his samples. This approach could seem risky at first, but results have demonstrated its efficacy.

As a compensation, in the las update, IMPUTE2 incorporated a feature that allows to add a second reference population to the imputation process (for example, another sample from the same population of the sample to impute), to help with the prediction. Conversely, other programs let the user decide which reference population wants to use for imputation (i.e. Europeans, CEU).

The issue with the reference populations is that they may be not that representative. As already mentioned, imputation appeared to work with common variants in the principal main population groups. That excludes rare variants and admixture populations. Moreover, the sample sizes of the reference populations were, till now, quite limited. For example, the previous HapMap project release 2 reference population did not surpass 300 samples and contained individuals from 5 different populations (Europeans are only 90 samples); it seems quite naïve to think that those haplotypes and recombination rates are actually representative of the entire population when the average GWAs has a bigger sample size.

In addition, reference populations are based on common markers. Given that only variants that are in the reference haplotypes will be imputed, if the sample to impute is enriched in rare variation, that information will not be considered.

Recently, other reference populations have been added to the imputation scheme (Michigan Imputation Server (Das et al., 2016)). They are characterized for having much larger number of samples and ethnic and phenotypic diversity. Current available reference populations will be briefly described next:

- HapMap Project (Belmont et al., 2005) release 3: 1115 samples. The main goal of the HapMap project was the creation of a haplotype map of the human genome (the common genetic variants catalog). The third release of data of this project contains samples from different 11 population groups: African ancestry in Southwest USA (ASW), Utah residents with Northern and Western European ancestry from the CEPH collection (CEU), Han Chinese in Beijing, China (CHB), Chinese in Metropolitan Denver, Colorado (CHD), Gujarati Indians in Houston, Texas (GIH), Japanese in Tokyo, Japan (JPT), Luhya in Webuye, Kenya (LWK), Mexican ancestry in Los Angeles, California (MXL), Maasai in Kinyawa, Kenya (MKK), Toscani in Italia (TSI), and Yoruba in Ibadan, Nigeria (YRI).

- 1KGP (Consortium, 2010): 1092 samples. This project was launched with the objective of finding as many rare genetic variants as possible (with frequency below 1%). Initially, reference population was based in a very small sample of ethnically homogeneous subjects. What is common for a population of reference can be rare in another. And this is especially relevant when working with underrepresented populations in reference panels (and/or high level of admixture). Phase 3 already incorporates 5,008 haplotypes from 26 populations across the world.

- Haplotype Reference Consortium (HRC): approximately 32500 samples of predominantly European ancestry; but eventually will also include the 1000 Genomes Project data.

- Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA): whole genome sequences were available on 883 individuals of African ancestry.

- TOPMed (TOPMed Freeze5 on GRCh38): it will consist of 25,568 haplotypes, mostly non-European. It is part of a broader Precision Medicine Initiative, which aims to provide disease treatments tailored to an individual's unique genes and

environment. Approximately 60% of the participants with substantial non-European ancestry.

4.1.3. Rationale

In the context of application geostatistical interpolation techniques in the purely genetic context, and seeing the good results obtained at the European scale, we decided to properly investigate if kriging can predict as good as conventional genetic imputation techniques, especially for low frequency markers.

In a previous comparative pilot study in our group (Lema Casal, 2015), results pointed towards a better performance of kriging when compared to imputation with MACH under less favorable genetic imputation conditions: low frequency variants (minor allele frequency (MAF)<10%) in a window with low density of markers. Results in Table 7 indicate the RMSE for the imputation was greater.

MAF<10% (N=59)	GENETIC IMPUTATION	KRIGING
RMSE	0.0228	0.0182

Table 7. RMSE calculated for N=75 rare variants of chromosome 22 when applied kriging and genetic imputation (MACH).

Moreover, rare variation is more connected to geography (Bycroft et al., 2019), so maybe spatially closer populations can predict better for rare variants than genetic imputation with reference population.

4.2. MATERIALS AND METHODS

4.2.1. Data and methods overview

For this application, data of chromosome 22 from 1172 anonymized Spanish samples were used. They had been genotyped with

the Genome-Wide Human SNP Array 6.0 chip from Affymetrix© and distributed in 32 spatial clusters, as Figure 23 indicates. Genotyped data were already published and available upon request for academic research use (Fernandez-Rozadilla et al., 2013; Bycroft et al., 2019).

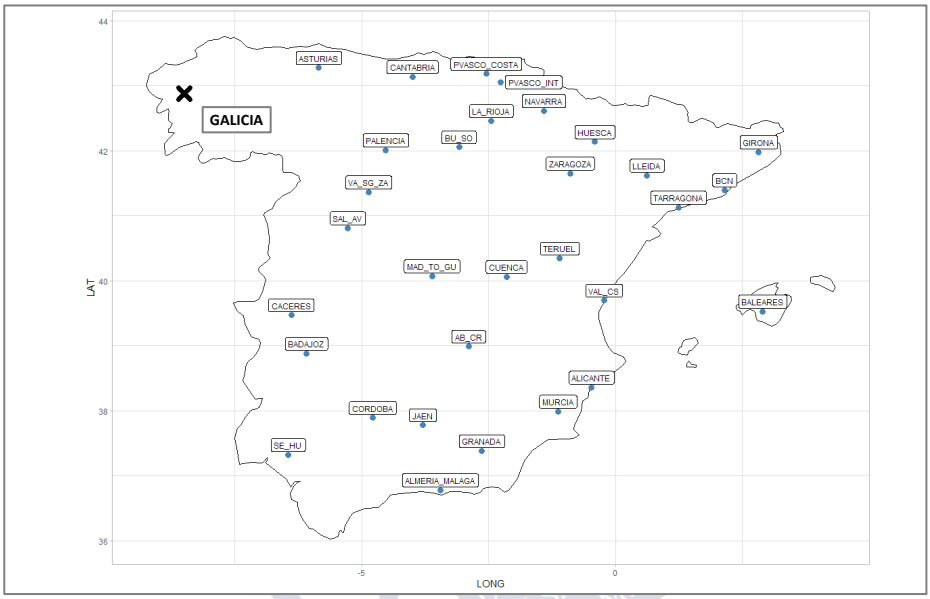

Figure 23. Map of the spatial clusters created for the analysis. Abbreviated labels of the clusters in alphabetical order: BCN: Barcelona, BU_SO: Burgos-Soria, MAD_TO_GU: Madrid-Toledo-Guadalajara, PVASCO_COSTA: Guipúzcoa-Vizcaya, PVASCO_INT: Álava, SAL_AV: Salamanca-Ávila, SE_HU: Sevilla-Huelva, VAL_CS: Valencia-Castellón, VA_SG_ZA: Valladolid-Segovia-Zamora.

Having selected the Galician cluster as a testing set (provided it is the largest cluster, the other ranging from population samples of 10 to up to 62 subjects), we sequentially removed a single SNP of the chromosome 22 and performed its genetic imputation in Galician samples. On the other hand, we used the allele frequency of that SNP in the remaining 31 spatial clusters to interpolate its value in Galician population with kriging. This data generation workflow is illustrated in Figure 24 and was carried out through custom in-house scripts.

In the end, we obtained two predicted values of frequency for each considered SNP in Galician population (one imputed, one kriged) that

were used to assess the performance of both techniques, when compared with the real allele frequency of the marker.

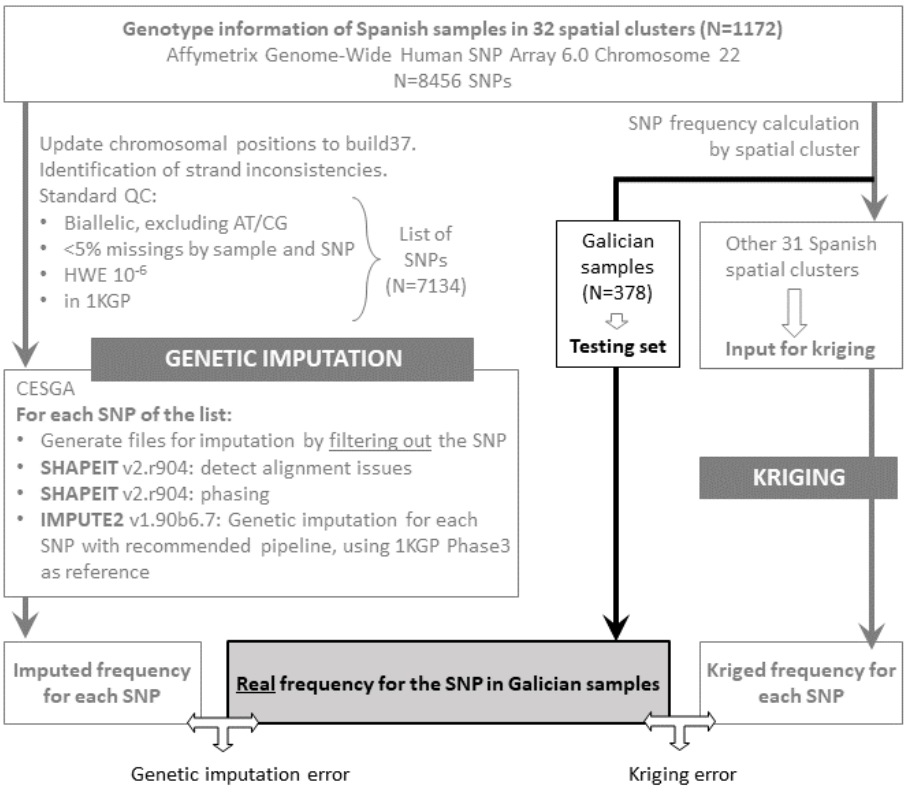

Figure 24. Workflow diagram of the data preparation and analysis.

Details for genetic imputation and kriging will be described in the next sections.

4.2.2. Genetic imputation

After renaming and updating chromosomic positions to the reference genome build GRCh37, a standard quality control (QC) of data was carried out in Plink v1.90b6.7 (Purcell et al., 2007) and R (R Foundation for Statistical Computing, 2018).

We followed a standard QC pipeline (Marees et al., 2018) except for filtering out by minor allele frequency, as we wanted to explore performance based on that characteristic. Only samples and SNPs that matched the following criteria were kept: biallelic, non-A/T-C/G markers with missing genotype rate <5% and that do not deviate from HWE; and samples with missing rate <5% (none of the samples was filtered out for this reason). Only SNPs present in 1KGP Phase 3 were maintained, otherwise they would not be imputed.

For each SNP of the QC-filtered list, chunks of the recommended size (buffer region of 250000bp at each side of the SNP position) were made.

A customized three-step shell script was designed and applied in an iterative manner, as Figure 24 briefly indicated:

1. Detection of alignment issues with SHAPEIT, comparing our data with 1KG Phase 3 Chromosome 22 reference genome. The script will automatically solve this using the output file from this first step.
2. Rerun SHAPEIT for phasing the Galician genotype samples excluding the target SNP.
3. Imputation of the target SNP using IMPUTE2 v.2.3.2 (B. N. Howie et al., 2009), following the recommended best practices: chunk[17] size of 5GB and effective population size (N_e)[18] $2e^6$.

This customized pipeline for imputation was accomplished in the FinisTerrae-II system of the Fundación Centro Tecnolóxico de Supercomputación de Galicia (CESGA), using a job array design to parallel up to 100 processes, with an assigned memory of 16GB and an

[17]Chunk: window for imputation, in our case, we set it as a buffer distance (in bp) at each flanking region of the target SNP.
[18]Effective population size (Ne) in the population-genetic model of IMPUTE2; in newer versions of the software it is already set by default to 2e6.

average run time for each marker of ~18min. Imputed output was obtained for 6599 SNPs with a quality score info metric[19]>0.4.

4.2.3. Kriging

For each spatial cluster, geographic coordinates (latitude and longitude) were manually assigned and each SNP frequency was calculated in Plink v1.90b6.7 (Purcell et al., 2007). The rescale of geographic coordinates to the range [0,1] was tested, with almost identical results in prediction; only non-rescaled analysis will be shown.

A customized in-house R script (R Foundation for Statistical Computing, 2018) was developed to automatically assess the spatial properties of each marker (isotropy, intrinsic stationarity, normality and spatial dependence), generate a matrix of eligible initial values of partial sill σ^2 and range ϕ for the minimization algorithm, and perform kriging (see Section 1.3.2.). Kriged output was obtained for N=8249 markers.

4.2.4. Analysis

Only SNPs that were successfully imputed and kriged (N=6084), with complete data will be considered for error calculation. All the analysis and graphical representations were carried out in R (R Foundation for Statistical Computing, 2018).

4.2.4.1. Error calculation

For each applied technique, several measurements of error were calculated, defined as follows:

[19]Info metric: metric about the quality of imputation (for more details, see Section 4.1).

- **Error:** difference between predicted (imputed or kriged) allele frequency and real allele frequency in our Galician population. It will inform if there is a predominant over- or underestimation.

- **Absolute error:** to assess the magnitude of error.

- **Relative error:** absolute error/real frequency; considers the real value of allele frequency, given that an error of 0.1 has not the same consequences over a high frequency of 0.6 than over a rare frequency of 0.03.

- **Root-mean square error (RMSE):** to compare both estimation techniques. As a less biases measure of error, it is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - observed_i)^2}{n}}$$

4.2.4.2. Kriging vs genetic imputation with IMPUTE2

Comparison of key features of each technique will be compiled in a table (see Section 4.3.1), such as input/output balance, computational time, possible bias factors, and different measures of error. Mean, minimum and maximum values were calculated. Spearman correlation was used to assess influence over relative error of these bias sources.

4.2.4.3. Influence of the allele frequency

SNPs will be classified in frequency categories and recalculation of errors by category will be carried out, to verify if the performance in different categories vary (it is known that rare variants perform worse with current genetic imputation methods).

4.3. RESULTS

4.3.1. Kriging vs genetic imputation with IMPUTE2

The following table summarizes the key features for the comparison:

	GENETIC IMPUTATION	KRIGING
Software Time (average per SNP)	SHAPEIT, IMPUTE2 ~18 min*	PLINK, R ~41 s*
Input files	Genotypes Recombination map Reference Genome	Allele frequencies at 31 locations
Input data (N)	7134	8456 in 31 locations
Output data (N)	6599**	8249
Possible bias	Chromosomal position, recombination rate, number of markers in the chunk, quality of imputation	Deviation from spatial assumptions, sampling bias
Perfect prediction	93	0
Good prediction***	1719	200
Error range Overestimated SNPs	±0.006 3117/6599	±0.02 3450/8249
Absolute error: magnitude	0.00250 (0 - 0.32340)	0.017146 ($6e^{-6}$ - 0.119477)
Relative error	0.010799 (0 - 1.334890)	1.334890 ($3e^{-5}$-1.2565917)
RMSE	0.009227015	0.02608943

Table 8. Overview of key features of genetic imputation and kriging. Absolute error and relative error are indicated as median (minimum value – maximum value)
*Time will vary depending on availability of resources; imputation was carried out in HPC environment, kriging in a PC RAM 8Gb
**post-imputation QC: info metric >0.4
***Good prediction: if absolute error <1e-3

The geostatistical approach (kriging) is faster and does not require a reference population to be carried out. In addition, a higher proportion of SNPs were successfully kriged (8249 out of the 8456, 97.6%) than imputed (6599 out of 8456, 78.0%).

Several factors could potentially influence results depending on the applied technique:

- For genetic imputation: chromosomal position, recombination rate, info metric, number of markers in the chunk used for imputation, and quality of imputation.

- For kriging, the compliance with the spatial properties (none of the markers met the four assumptions) and sampling bias.

No differences in errors were found based on chromosomal position ($rho_{Spearman}$=0.02, p= 0.0897), or number of markers included in the chunk ($rho_{Spearman}$=-0.01, p= 0.2665) for imputed values. Conversely, recombination rate ($rho_{Spearman}$=0.2, p=0) and quality of imputation (for info metric $rho_{Spearman}$=-0.61, p=0; for certainty $rho_{Spearman}$= -0.54, p=0) are correlated with relative error. This would imply that positions with high recombination rate or poor quality of imputation would have an increased relative error (worse performance), as expected.

Although surprisingly, there are several SNPs with a poor prediction, even though they have the right quality imputation metrics. These markers would not be detected in a standard post-imputation QC. The following table shows a sample of these markers, with an absolute error greater than 0.05.

SNP	Info metric	Certainty	Absolute error	Relative error
rs138711	0.999	0.999	0.3234	0.55
rs17653487	0.985	0.995	0.0729	0.65675676
rs4995261	0.519	0.712	0.0624	0.16165803
rs4822015	0.616	0.906	0.06003	0.57171429
rs1476035	0.652	0.778	0.0527	0.12227378
rs17608968	0.623	0.886	0.0515	0.39312977

Table 9. Sample of SNPs with absolute error>0.05.

IMPUTE2 outperforms kriging in prediction in terms of perfect or good prediction (absolute error<1e-3): 93 vs none; and 1719 vs 200, respectively.

Imputation errors are in the range ±0.006, whereas kriging errors are in the range ±0.02. Similar proportion of SNPs are overestimated in both techniques.

The following plot represents observed vs predicted values of allele frequency in our study (for markers properly predicted by the two approaches, N=6084). Simple visual comparison already points toward a higher dispersion of predicted values with kriging, as Figure 25 illustrates.

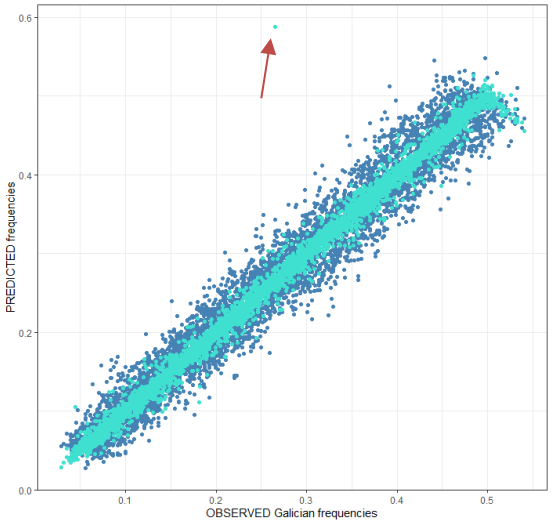

Figure 25. Correspondence of observed allele frequencies in Galician vs predicted by IMPUTE2 (in light blue) and kriging (in darker blue).

Anecdotally, there is an imputed marker that present the highest error of prediction (red arrow in Figure 25): rs138711 (C/T): real frequency=0.2646; imputed frequency=0.588. As previously commented, it was not filtered out in the post imputation QC.

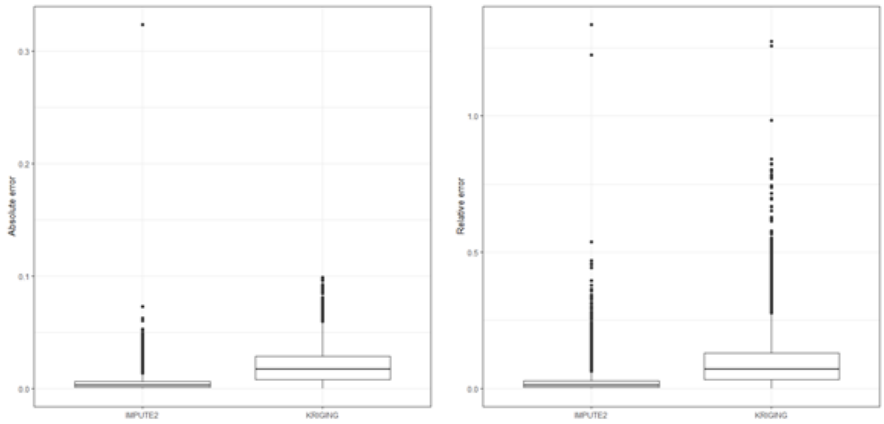

Figure 26. Boxplots of absolute and relative error for genetic imputation and kriging.

As illustrated in Figure 26, there are significant differences in absolute and relative error between genetic imputation and kriging (Wilcoxon $p<2e^{-16}$); the RMSE of kriging is greater (Table 8).

4.3.2. The influence of allele frequency: rare variants

Investigating the influence of allele frequency in the error of the predictions (of the markers properly predicted by the two approaches, N=6084), we divided SNPs in low ('rare'), low-intermediate, intermediate-high, and high frequency, and calculated error measure for each group. Results are shown in the next table:

MAF groups	Genetic imputation RMSE	Kriging RMSE
MAF <5% (N=67)	0.008719866	0.01746264
5%< MAF <15% (N=1434)	0.007338812	0.01973099
15%< MAF < 30% (N=2102)	0.01086337	0.02428053
MAF>30% (N=2481)	0.008875457	0.02801072

Table 10. RMSE by groups of MAF.

For kriging, RMSE is proportional to the allele frequency. Interestingly, for genetic imputation RMSE fluctuates with changing MAF, although not in a proportional manner: IMPUTE2 works best for low-intermediate, rare, and very common variants, in this order. Imputed SNPs of the intermediate-high category have RMSE levels close to kriging ones.

When plotting relative errors of genetic imputation and kriging against observed Galician frequencies, a major dispersion of values is found in the lower frequency range, as Figure 27 shows.

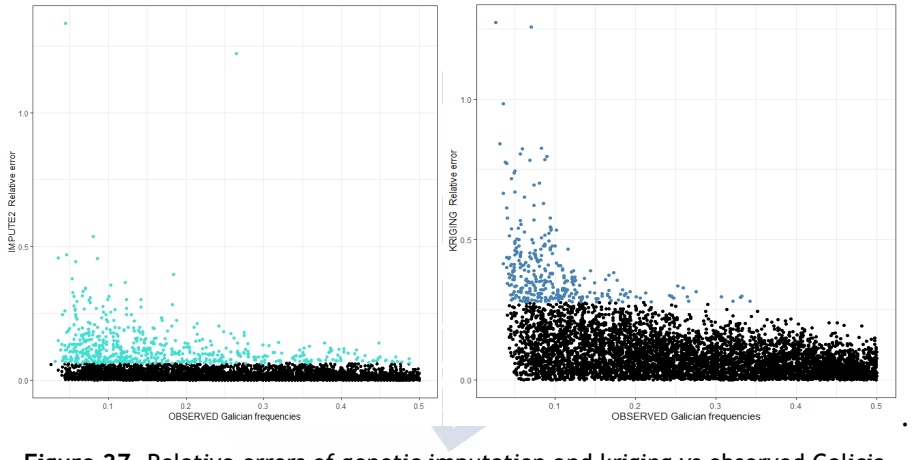

Figure 27. Relative errors of genetic imputation and kriging vs observed Galicia frequencies. Values of error out of range ± 1.5*IQR are colored.

RMSE calculation was repeated for the subset of 'rare' SNPs. Three different thresholds for 'rare' were set, due to the low number of data. Table 11 shows the results.

'Rare' variants	Genetic imputation RMSE	Kriging RMSE
MAF <5% (N=67)	0.008719866	0.01746264
MAF <10% (N=766)	0.006851057	0.01968972
MAF <15% (N=1501)	0.007405951	0.01963532

Table 11. RMSE for low frequency subsets of SNPs.

As before, kriging RMSE does not vary notably when restricting the analysis to low frequency SNPs, independently of the threshold set. On the contrary, genetic imputation RMSE is affected by the frequency of the markers considered.

Based on our previous work (Lema Casal, 2015), where density of the imputation window was definitive of the quality of imputation in MACH, we re-analyzed the possible effect of the number of markers included in the chunk joint to the selection of rare variants. The results indicate that the confluence of low frequency and low number of markers per chunk deteriorates imputation predictions to a level of error near the kriging errors.

'Rare' variants	RMSE imputation low density of markers	RMSE imputation high density of markers
MAF <5% (N=67)	0.01080799 (N=29)	0.004754452 (N=38)
MAF <10% (N=766)	0.00735652 (N=382)	0.006308167 (N=384)
MAF <15% (N=1501)	0.008176814 (N=739)	0.006572557 (N=762)

Table12. RMSE for low frequency subsets of SNPs considering number of markers per chunk.

5. DISCUSSION

Complex diseases result from the interplay of three contributors: genetics, environment, and their intertwined network of interactions (GxG, ExE, GxE). Nowadays, even with the current methodologies and advanced tools, the study of complex diseases remains challenging.

The present work summarizes the implementation of a geostatistical interpolation technique (kriging) in the study of complex diseases, first as a tool for integration of the effects of known genetic and environmental risk factors while exploring their relationships, and second, as an geostatistical interpolation technique for genetic allele frequencies, comparing its performance against conventional genetic imputation methods.

5.1. APPLICATION 1: MULTIPLE SCLEROSIS IN EUROPE[20]

MS is an autoimmune complex disease that affects more than 2.8 million people worldwide. MS displays a heterogenous geographical distribution, following a latitudinal gradient (higher prevalence in regions far from the Equator). Yet the primary cause of the disease remains uncertain, several genetic and environmental risk factors have

[20] To facilitate reading HLA-DRB1*15:01 will be referred as HLA-15; HLA-DRB1*03:01 as HLA-03; Pigmentation SNP: rs16891982 (G) as PIGM; VDR SNP: rs731236 (T) as VDR; and CYP27B1 SNP: rs12368653 (G) as CYP27B1 in the discussion.

been identified so far, some of them VD-related (see Section 3.1.3). Interestingly, MS patients have lower serum levels of VD than the general population, which postulates VD as a potential linker GxE, as seems to act as a connector among both genetic and environmental risk factors.

Limited by the modest number of prevalence data and the public availability of information about MS risk factors, we included HLA-15, HLA-03, PIGM, solar irradiation, VDR, and CYP27B1 in this study. To note that we worked under suboptimal conditions for kriging, with relatively few input data points of the predictors distributed unevenly across Europe.

Once data of risk factors were aligned to the MS prevalence locations through ordinary kriging, the correlation between predictors and MS prevalence was assessed. Our kriged data were consistent with evidence in the literature: positive correlation between HLA-15 and HLA-03 with MS (risk effect) and negative correlation between VDR and MS (protective effect).

Applying linear regression with a stepwise procedure of term selection resulted in a complex model with almost every interaction included, making it hard, if not impossible, to clearly understand the influence of each predictor over MS prevalence (K-R model). Thus, PCA was carried out before regression, to reduce dimensionality and resolve in part the collinearity issues that were present (K-PCA-R model).

Accounting for more than the 85% variance in the data, the first three PCs were selected. Their loadings are summarized as follows: PC1 accounts for almost half of the variation in the predictors and the predictors with higher loadings are HLA-15 (+0.54), solar irradiation (-0.53), and PIGM (+0.50); all of them with a marked latitudinal trend. PC2 is formed mostly by longitude (+0.58) and HLA-03 (-0.70). Finally, PC3 is almost entirely composed by VDR (-0.83), followed by longitude (+0.37) and PIGM (+0.36). In the associated biplots (see Figures 16-18) it can be observed that there is an almost perfect division

by European geographic region. Five clusters are easily defined, correspondingly: Nordic regions, East Mediterranean region (Italy), Central Europe, West Mediterranean region (Cyprus), and the British Islands.

The K-PCA-R model was after calculated and kept as predictors the three PCs and their 2-by-2 interactions. PC1 shows a positive relationship with MS prevalence. The relevance of solar irradiation and HLA-15 in this component corresponds to the known latitudinal gradient of MS: higher prevalence in Northern areas, where there are higher HLA-15 risk allele frequency and less solar irradiation. Although the role of both variables is integrated into the same PC, the results in the individual variables confirm that both are linked to MS prevalence. It should be noted that in PC1 the influence of PIGM reflects the well-established existing latitudinal cline of pigmentation.

The negative contribution of PC2 to MS indicates a W-E gradient not well documented, but that can be detected in the distribution of spatial clusters in the PC1-PC2 biplot (see Figure 16). It is worth mentioning that the allele HLA-03 was included in this study for being the principal risk factor linked to the higher prevalence in the Italian regions, specifically in Sardinia, with probably a microscale effect due to isolation.

Finally, PC3 has an overall protective contribution due to its highest loading (VDR), although counterbalanced in Eastern regions and/or populations with lighter skin.

The two significative interactions are of special interest: the positive interaction PC1:PC2 indicates an unexpected accentuation effect of PC1 in the East (in regions with low HLA-03), but always having in mind the principal effects observed in both PCs. On the other hand, the positive interaction between PC1:PC3 would potentiate PC1's effect in Northeastern regions.

Our results characterized a NW-SE gradient of MS prevalence, in which are involved both climatic variables (solar irradiation) and risk alleles (HLA) in combination with the protective effect of VDR.

The role of HLA alleles was already documented in numerous association studies in different populations (Sawcer et al., 2011; Mokry et al., 2016) and are considered the leaders of the high MS prevalence in determined areas, such as Sardinia (M. G. Marrosu et al., 2001).

The observed gradient in PIGM is thought to be originated by positive selection (Jablonski & Chaplin, 2000, 2012; Yuen & Jablonski, 2010) over lighter skins, that use more efficiently the solar irradiation in the endogenous synthesis of vitamin D. It would have been really interesting to confirm its role and interactions with the risk alleles and the solar irradiation but in the present study, the reduced variability between populations of the same latitude (in part due to the low number of included observations, N=30) did not allow us to.

In this first application of kriging, we were able to leverage available information from public databases and already published misaligned data of known risk factors of multiple sclerosis, into a model that reasonably describes the spatial distribution of the disease in Europe. As illustrated in the contour map of the K-PCA-R model (see Figure 17), kriged values tend to be smoother than the corresponding real ones (it is an intrinsic consequence of the kriging process), but that does not diminish the validity of our results. The spatial pattern of MS prevalence distribution in Europe was successfully mimicked with this approach, which demonstrated that this geostatistical interpolation technique is capable, even under suboptimal conditions, to capture the general spatial trends of MS prevalence in Europe.

5.2. APPLICATION 2: KRIGING VS GENETIC IMPUTATION

Previous work in our group pointed out the possibility that kriging could achieve better results predicting allele frequency than genetic

imputation when the imputation conditions were not optimal (low frequency of the SNP and low density of markers in the imputation window). It was decided to properly address the issue in a systematic way, using all the markers included in a complete chromosome 22.

An in-house custom pipeline with SHAPEIT and IMPUTE2 was implemented to run in an HPC environment (CESGA). Iteratively, each marker was removed from the original genotype file and imputed, with 1KGP chromosome 22 reference data.

In parallel, ordinary kriging was carried out using the other 31 geographical clusters as input. It should be noted that the computational cost of the genetic imputation is extremely high compared with the geostatistical approach.

Results from both prediction techniques were compared with real values of frequency of the Galician subpopulation (our testing set, N=378) and several measures of error were calculated. The final comparison was carried out with RMSE, also taking into consideration possible bias for each technique.

For genetic imputation, a general GWAS pipeline was applied, with pre and post imputation QC, without filtering for MAF, given our objective for this study (Marees et al., 2018). Still, and maybe anecdotally, the marker with the biggest error would have passed post imputation QC undetected, because quality scores for imputation were nearly perfect. It should be noted that the conditions of kriging were by far unfavorable; as opposed to the imputation ones, that were set using the recommended default. To note, the pre-imputation QC and the dependence of a reference population (only the markers present in the reference population will be imputed) restricts the output significantly. About the possible bias of imputation results, neither chromosome position nor number of markers in the chunk would directly correlate with increased error but, as expected, the recombination rate and quality metrics will.

Genetic imputation outperformed kriging in all groups of frequency across markers, even in the rare variants, in which we previously had found that the spatial approach could perform better (Lema Casal, 2015). Although this former work was based in a limited sample of 75 markers, the number of rare variants did not differ from the current number (N=59 before, N=67 now), implying that the inconsistency will surely be related to the updated software and reference population databases.

Comparing the performance at different frequency ranges, kriging had a much steady level of RMSE than IMPUTE2. This suggests that genetic imputation performance is influenced by frequency, and more severely if the number of markers included in the chunk is low. In this study, the lowest genetic imputation performance was achieved for SNPs with MAF<5% and low-density chunks.

In the case of kriging, regarding the spatial properties that guarantee the best kriging performance, none of the SNPs complied with all of them. Again, kriging was applied under non-optimal conditions. Ideally, input data for kriging will be a grid with plenty of informative points, based on samples with a minimum of observations to properly calculate allele frequency. In our case, clusters were limited in size and unevenly distributed. Giving the reduced availability of data, and to guarantee a minimum number of input points, some of the clusters were made with a modest number of subjects (ranging from 10 to 62 samples per cluster). This could condition the representability of the allele frequency calculated for some SNPs in some clusters, as well as the level of detection of results: for example, a frequency calculated from a spatial cluster of 20 individuals will at most, be of 1 count of minor allele over 2*20 total counts.

Likewise, the size of our target geographic population (in this case, one order of magnitude superior, N=378) will also condition the performance, as the error will be calculated over that frequency. Besides, our target point location is out of the spatial area of input data, making clear that we compared extreme scenarios (best conditions for genetic imputation vs worst conditions for kriging).

Even though the geographic scale for this second application is reduced when compared with the previous one (MS in Europe), it is well known that geographical patterns of genetic variability exist in Spain, especially in the North, between the populations from País Vasco and Galicia (Bycroft et al., 2019). This variability is evident by conducting PCA of a subset of independent SNPs, even when analyzing individual chromosomes. However, an additional exploration made with markers of chromosome 1 (where those differences were certainly present) did not comply with kriging requirements, meaning that either this subsamples are not enough representative to capture local geographic variability or that it is needed any more sophisticated geographical technique to work at this scale.

Future studies will ensure a greater number of samples (control population from BNADN) to properly assess the utility of kriging in optimal conditions and evaluate its applicability specifically in the case of rare variants.

6. CONCLUSIONS

1. The application of geostatistical interpolation techniques (kriging) to data of known genetic and environmental risk factors of multiple sclerosis manages to properly reproduce its heterogeneous spatial distribution of multiple sclerosis (MS) in Europe.

1.1. MS prevalence follows a NW-SE gradient, that is correctly reproduced by the K-PCA-R model.

1.2. The latitudinal gradient in MS is explained for the most part by the confluence of solar irradiation, HLA-DRB1*15:01 risk allele and the allele for lighter skin rs16891982.

1.3. The results point towards a modulator role of VDR but do not allow to correctly address the pigmentation effect.

2. Kriging is a flexible spatial interpolation technique, that can be combined with regular linear regression, reduction of dimensionality techniques and model selection methods, and performs well even under sub-optimal conditions.

3. Kriging does not surpass current genetic imputation methods (IMPUTE2) in the prediction of allele frequencies for markers in chromosome 22.

4. Genetic imputation with IMPUTE2 could perform worse with rare variants in low density of markers chromosomal regions.

7. RESUMO

Kriging: aplicando técnicas xeostatísticas no estudo xenético de enfermidades complexas.

As enfermidades complexas (aquelas nas que tanto factores xenéticos coma factores ambientais están implicados) posúen unha arquitectura que dificulta o seu estudo, dada a intricada rede de interaccións que teñen lugar (xene-xene, xene-ambiente, ambiente-ambiente) e ao coñecemento limitado sobre o momento no que esas interaccións son decisivas para a enfermidade.

En moitos casos a propia definición e características fenotípicas da doenza non están perfectamente descritas, estando os seus criterios diagnósticos en constante proceso de revisión e presentando signos e síntomas en espectro (con elevada variabilidade de presentación entre casos afectados), o cal lastra o deseño de estudos e condiciona a investigación no seu eido.

Por unha banda, o estudo xenético das enfermidades complexas vese limitado pola dificultade de asignar relevancia clínica aos achados xenéticos, e ao feito de que a realidade fisiopatolóxica é dinámica (non estática) e está modulada polo menos en parte por mecanismos epixenéticos, resultado tanto da herdanza como de exposicións ambientais. De xeito más xeneralizado e dende un punto de vista pragmático, a esta dificultade engádese moitas veces que é inviable a recollida de mostras do tecido afectado (cos condicionantes da correlación mostra-tecido enfermo), as periódicas actualizacións no xenoma de referencia e a dependencia de poboacións de referencia baseadas nun número limitado de mostras á hora de contrastar os

achados xenéticos, que reducen a potencia e reproducibilidade dos resultados. Todo isto complícase no caso de factores xenéticos de baixa frecuencia (variantes raras), posto que as técnicas e metodoloxías actuais están deseñadas para traballar con variación común. É por iso que cómpre o desenvolvemento de técnicas e metodoloxías de análise que permitan analizar correctamente este tipo de variación, que nestes momentos está tendo grande relevancia.

Por outra banda, en canto á investigación dos factores ambientais, a súa inclusión nos estudos de enfermidades complexas quedaba de xeito habitual reducida á adición de variables de axuste no modelo da enfermidade. Os datos sobre exposición ambiental adoitaban basearse na recollida de cuestionarios epidemiolóxicos de maneira retrospectiva ou prospectiva, moitas veces cubertos polo propio suxeito. En xeral, esta información ten comprometidas a súa consistencia e fiabilidade, ao non medir de forma estandarizada a exposición a factores ambientais de risco e a depender da memoria subxectiva (sendo imposible, polo tanto, obter información de exposicións no período embrionario, por exemplo; ou moi complicado facer un seguimento lonxitudinal prolongado, debido aos custos en tempo e recursos).

Na actualidade, o estudo sistemático de factores ambientais implicados na enfermidade está en grande auxe. Existen iniciativas de investigación en enfermidades complexas baseadas nun enfoque holístico: á par da busca de factores puramente xenéticos ligados á doenza, a recollida de datos amplíase a unha grande batería de elementos potencialmente informativos (estudos de biomarcadores, monitorización de parámetros sanguíneos, probas de imaxe, cuestionarios regulares de estilo de vida, dieta, exposicións ambientais e tratamentos farmacolóxicos) durante un amplo período de seguimento. Crese que a compoñente ambiental garda parte das respostas que poden explicar non só a orixe e desenvolvemento de patoloxías, senón tamén completar o coñecemento sobre os procesos e mecanismos que rexen a fisioloxía en condicións normais.

No caso das interaccións, non existe unha aproximación única xeneralizada, senón que vén determinada polo acceso a datos de

calidade e a recursos computacionais. Cando ambos están garantizados, as técnicas de *machine learning* son as preferidas, pese a que teñen unha interpretabilidade e traslación directa á práctica clínica reducidas. Nos outros casos, dependendo da existencia previa de hipóteses sobre que factores poden estar implicados na interacción ou o tipo de interacción observada (aditiva ou non), as metodoloxías de análise son variadas.

A compoñente xeográfica dos procesos biolóxicos foi recentemente establecida como unha base versátil sobre a cal integrar todo tipo de información xeo-referenciada, nacendo o campo da epidemioloxía xenética ecoxeográfica. A súa aplicación nas ciencias biomédicas deixou xa resultados de éxito, coma o seguimento de brotes virais, o estudo dos efectos do cambio climático nas enfermidades dependentes de vectores de transmisión, o mapeado da distribución xeográfica de enfermidades infecciosas, a implicación da polución do aire nos cadros de asma agudos, ou a identificación de agrupacións espaciais de casos en enfermidades complexas coma o cancro, a esclerose múltiple ou a artrite reumatoide.

Neste marco multidisciplinar no estudo de doenzas complexas, destaca a potencial utilidade dunha aproximación xeostatísttica para combinar información de distintas fontes a través da súa dimensión espacial. Aplicada neste caso, permitiría identificar posibles novos factores de risco ou factores protectores que puideran axudar a ter una visión global máis axustada da súa arquitectura, de maneira particular, naquelas doenzas que presentan patróns característicos de distribución espacial.

En xeral, as técnicas xeostatísticas ofrecen unha serie de métodos xa implementados para detectar e analizar patróns xeográficos e para facer inferencias en base aos datos da mostra. Unha das máis comunmente empregadas é o *kriging*. Trátase dunha técnica xeostatística de interpolación espacial que asume que un proceso espacial se pode resumir na suma dunha media constante e un proceso intrinsecamente estacionario, cun semivariograma coñecido. O termo interpolación fai referencia á estimación dunha variable nunha localización non medida a partir de valores observados en localizacións

próximas a esa. Os algoritmos de interpolación estiman o valor na localización dada como unha suma ponderada dos valores da variable nas localizacións adxacentes, de acordo con funcións que outorgan un peso decrecente a medida que a distancia de separación aumenta. O *kriging* pode entenderse como un proceso en dous pasos: primeiro, determínase a estrutura da covarianza espacial nos puntos coñecidos mediante o axuste dun (semi)variograma; e segundo, os pesos derivados desa estrutura de covarianza son empregados para interpolar os valores da variable nas localizacións non coñecidas.

O presente traballo resume a implementación dunha técnica de interpolación espacial (o *kriging*) no estudo de enfermidades complexas, primeiro como una ferramenta de integración dos efectos de varios factores de risco e protectores coñecidos de esclerose múltiple; e segundo, como una técnica de interpolación espacial de frecuencias alélicas, contrastando o seu desempeño co de metodoloxías de imputación xenética convencionais.

Para iso, definíronse os seguintes obxectivos: (a) Avaliar a aplicación dunha técnica xeostatística de interpolación (*kriging*) no estudo de enfermidades complexas cunha distribución xeográfica heteroxénea, como ferramenta para a integración de información procedente de distintas fontes e a distinta resolución espacial para reproducir o seu patrón espacial: tomando como enfermidade de estudo a esclerose múltiple, caracterizarase o seu gradiente xeográfico en Europa e analizarase a importancia relativa dos factores de risco e protectores relacionados coa vitamina D incluídos no estudo; e (b) Aplicar *kriging* como una alternativa aos métodos de imputación xenética convencionais, con énfase no rendemento baixo varios factores de influencia.

(a) Aplicación 1: Esclerose múltiple en Europa.

A esclerose múltiple é unha enfermidade complexa autoinmune do sistema nervioso central, que afecta a máis de 2.8 millóns de persoas en todo o mundo. Posúe una distribución xeográfica heteroxénea, seguindo

un gradiente latitudinal (maior prevalencia de esclerose múltiple en rexións máis afastadas do Ecuador).

Aínda que a causa orixinaria da doenza segue a ser descoñecida, varios factores de risco xenéticos e ambientais xa foron descubertos e descritos: o xénero feminino (ratio 3:1 en España), historia familiar de casos da enfermidade, ancestralidade europea, certas variantes xenéticas, historia previa de mononucleosis infecciosa, hábito tabáquico, obesidade e dieta, entre outros. Curiosamente, os afectados de esclerose múltiple posúen uns niveis séricos de vitamina D inferiores aos da poboación xeral. Isto, ligado á distribución latitudinal da enfermidade (que coincide inversamente coa distribución de irradiación solar, necesaria para a produción endóxena de vitamina D), axudou a postular á vitamina D como conector entre xenotipo e ambiente nesta enfermidade complexa. Baseándonos nesa hipótese, establecemos a hipótese de traballo para a primeira aplicación do *kriging*. Para a recompilación de datos, priorizamos factores relacionados coa enfermidade que seguirán os seguintes criterios (de aquí en diante, referidos como 'preditores'): factores relacionados coa vitamina D, cunha distribución espacial característica e establecida, relevantes en poboación europea e de acceso doado e gratuíto, ben recollidos en bases de datos públicas, ou dispoñibles en literatura xa publicada.

Seleccionáronse seis factores relacionados coa esclerose múltiple e/ou coa vitamina D: dous alelos HLA directamente ligados cun maior risco da doenza en poboación de ascendencia europea (HLA-DRB1*15:01 e HLA-DRB1*03:01, este último ligado especialmente a poboación italiana), a irradiación solar, como compoñente ambiental imprescindible para a produción endóxena de vitamina D, un marcador de ancestralidade ligado a ascendencia europea e pel clara, que sintetiza vitamina D de xeito máis eficiente (rs16891982), un marcador do receptor da vitamina D (VDR), último actor necesario para que a vitamina D acade o núcleo celular e actúe como modulador da transcrición (alelo ligado a efecto protector sobre a enfermidade), e un SNP (*single nucleotide polymorphism*) asociado a CYP27B1, o enzima responsable da bioactivación da vitamina D.

Limitados polo número modesto de datos de partida de prevalencia de esclerose múltiple e dos factores de risco e protectores considerados, realizouse *kriging* sobre os datos de cada preditor, aliñándoos coas localizacións nas que estaban dispoñibles os datos de prevalencia. Una vez comprobado que as correlacións dos valores krigeados dos preditores coa prevalencia non contradicían o coñecemento actual (os alelos HLA correlacionan positivamente coa prevalencia (efecto de risco) e VDR correlaciona negativamente (efecto protector)), aplicouse regresión lineal cunha selección de termos baseada no Akaike Information Criterion (AIC), considerando todas as posibles interaccións dúas a dúas. O modelo resultante era complicado de interpretar polo que se decidiu engadir un paso previo de redución de dimensións mediante análise de compoñentes principais. Seleccionáronse os tres primeiros compoñentes principais (PC), que explican máis do 85% da variabilidade dos datos, e repetiuse o proceso de selección de termos na regresión lineal. O primeiro PC está composto por HLA-DRB1*15:01, irradiación solar e rs16891982; o segundo PC por lonxitude e HLA-DRB1*03:01; e finalmente, o terceiro está practicamente composto por VDR. Así pois, a redución de dimensións recolle esencialmente os factores de risco dos que hai evidencia de influencia sobre a enfermidade. A representación gráfica do modelo seleccionado (modelo K-PCA-R) reproduce fielmente a distribución xeográfica real da prevalencia de esclerose múltiple en Europa, definindo un gradiente noroeste-surleste e sendo o suficiente preciso como para ter en conta a as zonas de alta prevalencia que romperían o gradiente latitudinal orixinal (como sería o caso de Sardeña).

A primeira aplicación do *kriging* permitiu combinar información de factores de risco e protectores relacionados coa vitamina D en esclerose múltiple procedentes de bases de datos de acceso público e datos xa publicados referidos a distintas localizacións xeográficas e aliñalos para poder construír un modelo de prevalencia da enfermidade que de xeito aceptable reflexa a distribución especial heteroxénea desta doenza complexa en Europa.

(b) Aplicación 2: *Kriging* fronte a imputación xenética convencional.

A imputación xenética é a día de hoxe un paso máis na secuencia de procesos dos estudos de asociación xenética e en meta-análises, debido á composición variable de marcadores nos *arrays* comerciais de xenotipado. Na actualidade, existen varios programas de imputación xenética dispoñibles, sendo SHAPEIT-IMPUTE2, MACH-minimac e BEAGLE algúns dos exemplos máis comunmente empregados.

Aínda que cada un posúe as súas propias particularidades, todos comparten unha base conceptual común, que é o uso de modelos de Markov ocultos (Hidden Markov Models) para estimar haplotipos antes da imputación, tendo en conta o desequilibrio de ligamento e a taxa de recombinación das rexións da mostra a imputar e o xenotipo de referencia.

Na segunda aplicación de técnicas xeostatísticas dentro dun contexto puramente xenético, decidiuse investigar o desempeño do *kriging* como método de predición de frecuencias alélicas, contrapoñéndoo aos métodos de imputación xenética convencionais (neste caso SHAPEIT-IMPUTE2).

Previamente no noso grupo, nun estudo a pequena escala, veuse que esta técnica de interpolación podería ser una alternativa válida á imputación convencional, cun menor tempo de execución e resultados comparables en marcadores de baixa frecuencia (frecuencia do alelo menor <10%) e baixa densidade de marcdores na ventá de imputación. Por iso, empregando datos xa publicados de poboación española (N=1172), testouse esta vez a totalidade do cromosoma 22 para comprobar esta hipótese.

Mediante o deseño personalizado dunha secuencia de traballo propia, posta a punto en R de xeito local e no entorno high-performance computing (HPC) do CESGA, procedeuse a aplicar ambas metodoloxías de predición, *kriging* e imputación xenética, en paralelo para cada un dos marcadores do cromosoma 22.

Respecto ao *kriging*, a mostra de poboación española asignouse a 32 grupos xeográficos, dándolle coordenadas xeográficas e calculando frecuencias de alélicas para cada SNP en cada grupo especial. Seleccionouse o grupo asignado a Galicia como grupo de avaliación do desempeño das técnicas (*testing set*), dado que se trataba do grupo maioritario (N=378), cun límite de detección de erro maior. As frecuencias calculadas en cada un dos 31 puntos xeográficos restantes utilizaranse como información de partida para o *kriging* que, de entrada ten un menor tempo de computación que a imputación xenética.

Respecto á imputación xenética con IMPUTE2, aplicouse un protocolo estándar de control de calidade pre e post imputación, e as condicións de imputación foron as recomendadas (óptimas). Con todo, esta aproximación vese condicionada por varias limitacións: só se imputaron marcadores presentes na poboación de referencia, que pasasen os controis de calidade pre e post-imputación e cun elevado tempo de computación. A secuencia de procesos de traballo ocorre en dous pasos: *phasing* e detección de problemas de aliñamento, seguido de imputación con IMPUTE2, cunha media de 18 minutos de tempo por marcador.

Ao final obtivéronse dúas frecuencias alélicas preditas por SNP, unha de cada aproximación. Comparáronse coa frecuencia real no grupo xeográfico de galegos, calculando catro medidas de erro: erro, erro en valor absoluto, erro relativo á frecuencia, e erro cuadrático promedio (RMSE, polas súas siglas en inglés). Tivéronse en conta posibles factores que influenciaran o resultado das predicións. Na imputación xenética, a posición cromosómica, o mapa de recombinación e o número de marcadores incluídos na ventá de imputación. No *kriging*, o cumprimento coas propiedades espaciais testadas (aínda que non de xeito estrito, pois non parece afectar sensiblemente aos resultados). E finalmente, nos dous casos, avaliouse o erro por rangos de frecuencia do SNP e densidade de marcadores na ventá de imputación, con especial interese debido aos resultados previos en variantes raras.

Comparando o desempeño de ambas aproximacións en diferentes rangos de frecuencia alélica, o *kriging* tivo un nivel de erro máis alto que IMPUTE2, pero máis consistente e proporcional á frecuencia. Cando se avaliou a influencia da densidade de marcadores na ventá de imputación, comprobouse que IMPUTE2 empeora o seu desempeño en ventás de imputación con un número máis baixo de marcadores. Neste estudo, o peor desempeño da imputación xenética foi obtido para SNPs cunha frecuencia alélica <5% e ventás de imputación de con baixa densidade de marcadores.

No caso do *kriging*, considerando as propiedades espaciais que garantirían un desempeño óptimo da técnica, ningún dos marcadores considerados na mostra española cumpren as condición de isotropía, estacionaridade intrínseca, normalidade e dependencia espacial. Ademais, a mostra de partida conta con relativamente poucos puntos xeográficos repartidos de xeito irregular, e o punto a predicir atópase fóra da nube de puntos inicial. As limitación neste caso, ademais, aplícanse ao propio cálculo de frecuencias alélicas de partida. A limitada dispoñibilidade de datos fai que algunhas frecuencias foran calculadas sobre un número reducido de individuos (N<20), podendo afectar á representatividade dese dato de frecuencia a nivel da súa área espacial asignada.

Aínda que a escala xeográfica para esta segunda aplicación é reducida se a comparamos coa da Aplicación 1, é coñecido que en España existen patróns xeográficos de variabilidade xenética, especialmente nas rexións do norte, entre poboacións do País Vasco e Galicia. Esta variabilidade é evidente levando a cabo unha análise de compoñentes principais nunha submostra de SNPs independentes, mesmo analizando cromosomas individuais. Nunha proba feita cos marcadores do cromosoma 1, onde esas diferenzas estaban presentes, os requirimentos das propiedades espaciais para un desempeño óptimo do *kriging* non se acadaban tampouco, implicando que ou ben as submostras poboacionais non son representativas abondo para capturar a variabilidade xeográfica local ou que se necesita una adaptación da técnica de interpolación espacial para traballar a esa escala local reducida. Estudos futuros asegurarán un maior número de mostras (por

exemplo, a poboación control do Banco Nacional de ADN) para estudar adecuadamente esta hipótese e avaliar correctamente a súa aplicabilidade no caso específico de variantes de baixa frecuencia.

Deberanse ter en conta as limitacións deste estudo: a dependencia de datos de libre acceso, cun modesto tamaño de mostra e unha distribución xeográfica irregular, condicionou o desempeño do *kriging*, que en ambas aplicacións foi empregado en condicións non óptimas.

Con todo, o traballo realizado permitiu chegar ás seguintes conclusións:

En primeiro lugar, a aplicación de técnicas xeostatísticas de interpolación espacial (*kriging*) a datos de factores de risco xenéticos e ambientais coñecidos de esclerose múltiple consegue reproducir correctamente a súa distribución espacial heteroxénea en Europa. Isto permitiu identificar un gradiente noroeste-surleste da prevalencia desta enfermidade en Europa, que é correctamente reproducido polo modelo K-PCA-R (*krig – principal component analysis – regress*). Ademais, o gradiente latitudinal observado en esclerose múltiple está explicado na súa maior parte pola confluencia da irradiación solar, o alelo de risco HLA-DRB1*15:01 e o alelo para pel clara do SNP rs16891982. Os resultados obtidos indican un posible papel modulador do receptor de vitamina D (VDR), pero non permiten estudar adecuadamente o efecto, aparentemente contraditorio, da pigmentación (rs16891982) sobre o risco da enfermidade. Empregando a esclerose múltiple como caso de estudo, o *kriging* demostrou ser unha ferramenta flexible e valiosa, que pode ser combinada con técnicas de regresión lineal, de redución de dimensións e de métodos de selección de variables, para a integración de información procedente de varias fontes e a distinta resolución espacial nun modelo que permite visualizar doadamente a súa distribución heteroxénea en Europa e explorar as complexas interaccións entre varios dos seus factores de risco xenéticos e ambientais xa coñecidos.

En segundo lugar, na avaliación do kriging como alternativa ás técnicas de imputación xenética convencionais, pese a que o *kriging* non mellora

os resultados obtidos con IMPUTE2 na predición de frecuencias alélicas para marcadores do cromosoma 22, este estudo piloto puxo de manifesto o peor rendemento destas últimas para variantes raras (de baixa frecuencia) en rexións cromosómicas con baixa densidade de marcadores.

8. REFERENCES

Abbas, S., Nieters, A., Linseisen, J., Slanger, T., Kropp, S., Mutschelknauss, E. J., Flesch-Janys, D., & Chang-Claude, J. (2008). Vitamin D receptor gene polymorphisms and haplotypes and postmenopausal breast cancer risk. *Breast Cancer Research*, *10*(2). https://doi.org/10.1186/bcr1994

Agliardi, C., Guerini, F. R., Saresella, M., Caputo, D., Leone, M. A., Zanzottera, M., Bolognesi, E., Marventano, I., Barizzone, N., Fasano, M. E., Al-Daghri, N., & Clerici, M. (2011). Vitamin D receptor (VDR) gene SNPs influence VDR expression and modulate protection from multiple sclerosis in HLA-DRB1*15-positive individuals. *Brain, Behavior, and Immunity*, *25*(7), 1460–1467. https://doi.org/10.1016/j.bbi.2011.05.015

Al-Temaimi, R., AbuBaker, J., Al-khairi, I., & Alroughani, R. (2017). Remyelination modulators in multiple sclerosis patients. *Experimental and Molecular Pathology*, *103*(3), 237–241. https://doi.org/10.1016/j.yexmp.2017.11.004

Alemany-Navarro, M., Costas, J., Real, E., Segalàs, C., Bertolín, S., Domènech, L., Rabionet, R., Carracedo, Á., Menchón, J. M., & Alonso, P. (2019). Do polygenic risk and stressful life events predict pharmacological treatment response in obsessive compulsive disorder? A gene–environment interaction approach. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0410-0

Alla, S., & Mason, D. F. (2014). Multiple sclerosis in New Zealand. *Journal of Clinical Neuroscience*, *21*(8), 1288–1291. https://doi.org/10.1016/j.jocn.2013.09.009

Alonso-Gonzalez, A., Calaza, M., Rodriguez-Fontenla, C., & Carracedo, A. (2019). Novel gene-based analysis of ASD GWAS: Insight into the biological role of associated genes. *Frontiers in Genetics*, *10*(JUL), 1–11. https://doi.org/10.3389/fgene.2019.00733

Antonopoulou, K., Stefanaki, I., Lill, C. M., Chatzinasiou, F., Kypreou, K. P., Karagianni, F., Athanasiadis, E., Spyrou, G. M., Ioannidis, J. P. A., Bertram, L., Evangelou, E., & Stratigos, A. J. (2015). Updated Field Synopsis and Systematic Meta-Analyses of Genetic Association Studies in Cutaneous Melanoma: The MelGene Database. *Journal of Investigative Dermatology*, *135*(4), 1074–1079. https://doi.org/10.1038/jid.2014.491

ANZ. (2009). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nature Genetics*, *41*(7), 824–828. https://doi.org/10.1038/ng.396

Ascherio, A., Munger, K. L., White, R., Köchert, K., Simon, K. C., Polman, C. H., Freedman, M. S., Hartung, H. P., Miller, D. H., Montalbán, X., Edan, G., Barkhof, F., Pleimes, D., Radü, E. W., Sandbrink, R., Kappos, L., & Pohl, C. (2014). Vitamin D as an early predictor of multiple sclerosis activity and progression. *JAMA Neurology*, *71*(3), 306–314. https://doi.org/10.1001/jamaneurol.2013.5993

Athanasiadis, E. I., Antonopoulou, K., Chatzinasiou, F., Lill, C. M., Bourdakou, M. M., Sakellariou, A., Kypreou, K., Stefanaki, I., Evangelou, E., Ioannidis, J. P. A., Bertram, L., Stratigos, A. J., & Spyrou, G. M. (2014). A Web-based database of genetic association studies in cutaneous melanoma enhanced with network-driven data exploration tools. *Database : The Journal of Biological Databases and Curation*, *2014*, 1–13. https://doi.org/10.1093/database/bau101

Avila, M., Bansal, A., Culberson, J., & Peiris, A. N. (2018). The Role of Sex Hormones in Multiple Sclerosis. *European Neurology*,

19430, 93–99. https://doi.org/10.1159/000494262

Baranzini, S. E., Mudge, J., Van Velkinburgh, J. C., Khankhanian, P., Khrebtukova, I., Miller, N. A., Zhang, L., Farmer, A. D., Bell, C. J., Kim, R. W., May, G. D., Woodward, J. E., Caillier, S. J., McElroy, J. P., Gomez, R., Pando, M. J., Clendenen, L. E., Ganusova, E. E., Schilkey, F. D., … Kingsmore, S. F. (2010). Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, *464*(7293), 1351–1356. https://doi.org/10.1038/nature08990

Baranzini, S. E., & Oksenberg, J. R. (2017). The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years. *Trends in Genetics*, *xx*, 1–11. https://doi.org/10.1016/j.tig.2017.09.004

Baranzini, S. E., Srinivasan, R., Khankhanian, P., Okuda, D. T., Nelson, S. J., Matthews, P. M., Hauser, S. L., Oksenberg, J. R., & Pelletier, D. (2010). Genetic variation influences glutamate concentrations in brains of patients with multiple sclerosis. *Brain*, *133*(9), 2603–2611. https://doi.org/10.1093/brain/awq192

Barroso, E., Fernandez, L. P., Milne, R. L., Pita, G., Sendagorta, E., Floristan, U., Feito, M., Aviles, J. A., Martin-Gonzalez, M., Arias, J. I., Zamora, P., Blanco, M., Lazaro, P., Benitez, J., & Ribas, G. (2008). Genetic analysis of the vitamin D receptor gene in two epithelial cancers: Melanoma and breast cancer case-control studies. *BMC Cancer*, *8*. https://doi.org/10.1186/1471-2407-8-385

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., & Deckmyn, A. (2018). *maps: Draw Geographical Maps*.

Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., Cotsapas, C., Shah, T. S., Spencer, C., Booth, D., Goris, A., Oturai, A., Saarela, J., Fontaine, B., Hemmer, B., Martin, C., Zipp, F., D'Alfonso, S., Martinelli-Boneschi, F., Taylor, B., … McCauley, J. L. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis.

Nature Genetics, *45*(11), 1353–1362. https://doi.org/10.1038/ng.2770

Belmont, J. W., Boudreau, A., Leal, S. M., Hardenbol, P., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., … Stewart, J. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Beretich, B. D., & Beretich, T. M. (2009). Explaining multiple sclerosis prevalence by ultraviolet exposure: A geospatial analysis. *Multiple Sclerosis*, *15*(8), 891–898. https://doi.org/10.1177/1352458509105579

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Arthur, L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Tarjei, S., Thomson, J. A., Bernstein, B. E., Meissner, A., Kellis, M., Lander, E. S., & Mikkelsen, T. S. (2013). The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol.*, *28*(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045.The

Bhargava, P., Steele, S. U., Waubant, E., Revirajan, R., Marcus, J., Dembele, M., Sandra, D., Hollis, B. W., Crainiceanu, C., Mowry, E. M., & Francisco, S. (2016). *Multiple sclerosis patients have a diminished serologic response to vitamin D supplementation compared to healthy controls*. *22*(6), 753–760. https://doi.org/10.1177/1352458515600248.Multiple

Bien, S. A., Wojcik, G. L., Hodonsky, C. J., Gignoux, C. R., Cheng, I., Matise, T. C., Peters, U., Kenny, E. E., & North, K. E. (2019). The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE. *Annual Review of Genomics and Human Genetics*, *20*(1), 181–200. https://doi.org/10.1146/annurev-genom-091416-035517

Bjørnevik, K., Chitnis, T., Ascherio, A., & Munger, K. L. (2017). Polyunsaturated fatty acids and the risk of multiple sclerosis. *Multiple Sclerosis*, *23*(14), 1830–1838. https://doi.org/10.1177/1352458517691150

Booth, D. R., Ding, N., Parnell, G. P., Shahijanian, F., Coulter, S., Schibeci, S. D., Atkins, A. R., Stewart, G. J., Evans, R. M., Downes, M., & Liddle, C. (2016). Cistromic and genetic evidence that the Vitamin D receptor mediates susceptibility to latitude-dependent autoimmune diseases. *Genes and Immunity*, *17*(4), 213–219. https://doi.org/10.1038/gene.2016.12

Boström, I., & Landtblom, A. M. (2015). Does the changing sex ratio of multiple sclerosis give opportunities for intervention? *Acta Neurologica Scandinavica*, *132*(S199), 42–45. https://doi.org/10.1111/ane.12430

Bowman, A. W., & Azzalini, A. (2018). *R package "sm": nonparametric smoothing methods*.

Bowman, Adrian W., & Crujeiras, R. M. (2013). Inference for variograms. *Computational Statistics and Data Analysis*, *66*, 19–31. https://doi.org/10.1016/j.csda.2013.02.027

Breuer, J., Schwab, N., Schneider-Hohendorf, T., Marziniak, M., Mohan, H., Bhatia, U., Gross, C. C., Clausen, B. E., Weishaupt, C., Luger, T. a., Meuth, S. G., Loser, K., & Wiendl, H. (2014). Ultraviolet B light attenuates the systemic immune response in central nervous system autoimmunity. *Annals of Neurology*, *75*, 739–758. https://doi.org/10.1002/ana.24165

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Brownlee, W. J., Hardy, T. A., Fazekas, F., & Miller, D. H. (2017). Diagnosis of multiple sclerosis: progress and challenges. *The Lancet*, *389*(10076), 1336–1346. https://doi.org/10.1016/S0140-6736(16)30959-X

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bycroft, C., Fernandez-Rozadilla, C., Ruiz-Ponte, C., Quintela, I., Carracedo, Á., Donnelly, P., & Myers, S. (2019). Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-018-08272-w

Carrel, M., & Emch, M. (2014). Genetics: A New Landscape for Medical Geography. *Ann Assoc Am Geogr*, *103*(6), 1452–1467. https://doi.org/10.1038/jid.2014.371

Cavalli-Sforza, L. L. (2005). The human genome diversity project: Past, present and future. *Nature Reviews Genetics*, *6*(4), 333–340. https://doi.org/10.1038/nrg1596

Chatzinasiou, F., Lill, C. M., Kypreou, K., Stefanaki, I., Nicolaou, V., Spyrou, G., Evangelou, E., Roehr, J. T., Kodela, E., Katsambas, A., Tsao, H., Ioannidis, J. P. A., Bertram, L., & Stratigos, A. J. (2011). Comprehensive field synopsis and systematic meta-analyses of genetic association studies in cutaneous melanoma. *Journal of the National Cancer Institute*, *103*(16), 1227–1235. https://doi.org/10.1093/jnci/djr219

Chen, J., Chia, N., Kalari, K. R., Yao, J. Z., Novotna, M., Soldan, M.

M. P., Luckey, D. H., Marietta, E. V., Jeraldo, P. R., Chen, X., Weinshenker, B. G., Rodriguez, M., Kantarci, O. H., Nelson, H., Murray, J. A., & Mangalam, A. K. (2016). Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. *Scientific Reports*, *6*, 1–10. https://doi.org/10.1038/srep28484

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., & Dreyer, K. J. (2018). Current applications and future impact of machine learning in radiology. *Radiology*, *288*(2), 318–328. https://doi.org/10.1148/radiol.2018171820

Christensen, T. (2017). Human endogenous retroviruses in the aetiology of MS. *Acta Neurologica Scandinavica*, *136*, 18–21. https://doi.org/10.1111/ane.12836

Colombini, A., Brayda-Bruno, M., Lombardi, G., Croiset, S. J., Ceriani, C., Buligan, C., Barbina, M., Banfi, G., & Cauci, S. (2016). BsmI, ApaI and TaqI polymorphisms in the Vitamin D Receptor gene (VDR) and association with lumbar spine pathologies: An Italian case-control study. *PLoS ONE*, *11*(5), e0155004. https://doi.org/10.1371/journal.pone.0155004

Colón-González, F. J., Fezzi, C., Lake, I. R., & Hunter, P. R. (2013). The Effects of Weather and Climate Change on Dengue. *PLoS Neglected Tropical Diseases*, *7*(11), 1–9. https://doi.org/10.1371/journal.pntd.0002503

Compston, A. (1997). Genetic epidemiology of multiple sclerosis. *Epidemiologic Reviews*, *19*(1), 99–106.

Consortium, T. 1000 G. P. (2010). A map of human genome variation from population-scale sequencing Accessed UKPMC Funders Group Author Manuscript. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534.A

Correale, J., Ysrraelit, M. C., & Gaitn, M. I. (2009). Immunomodulatory effects of Vitamin D in multiple sclerosis. *Brain*, *132*, 1146–1160. https://doi.org/10.1093/brain/awp033

Cree, B., Khan, O., Bourdette, D., Goodin, D., Cohen, J., Marrie, R., Glidden, D., Weinstock-Guttman, B., Reich, D., Patterson, N., Haines, J., Pericak-Vance, M., DeLoa, C., Oksenberg, J., & Hauser, S. (2004). Clinical characteristics of African Americans vs Caucasian Americans with multiple sclerosis. *Neurology*, *63*(11), 2039–2045. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed6&NEWS=N&AN=2004531204

Cressie, N. (1993). *Statistics for spatial data* (J. W. & Sons (ed.)).

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Degelman, M. L., & Herman, K. M. (2017). Smoking and multiple sclerosis: A systematic review and meta-analysis using the Bradford Hill criteria for causation. In *Multiple sclerosis and related disorders* (Vol. 17). Elsevier B.V. https://doi.org/10.1016/j.msard.2017.07.020

Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F., & Marchini, J. (2013). Haplotype estimation using sequencing reads. *American Journal of Human Genetics*, *93*(4), 687–696. https://doi.org/10.1016/j.ajhg.2013.09.002

Delaneau, O., Marchini, J., McVeanh, G. A., Donnelly, P., Lunter, G., Marchini, J. L., Myers, S., Gupta-Hinch, A., Iqbal, Z., Mathieson, I., Rimmer, A., Xifara, D. K., Kerasidou, A., Churchhouse, C.,

Altshuler, D. M., Gabriel, S. B., Lander, E. S., Gupta, N., Daly, M. J., … Peltonen, L. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, *5*, 1–9. https://doi.org/10.1038/ncomms4934

Delaneau, O., Marchini, J., & Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*(2), 179–181. https://doi.org/10.1038/nmeth.1785

Delaneau, O., Zagury, J. F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. https://doi.org/10.1038/nmeth.2307

Dendrou, C. A., Fugger, L., & Friese, M. A. (2015). Immunopathology of multiple sclerosis. *Nature Reviews Immunology*, *15*(9), 545–558. https://doi.org/10.1038/nri3871

Dhewantara, P. W., Lau, C. L., Allan, K. J., Hu, W., Zhang, W., Mamun, A. A., & Soares Magalhães, R. J. (2019). Spatial epidemiological approaches to inform leptospirosis surveillance and control: A systematic review and critical appraisal of methods. *Zoonoses and Public Health*, *66*(2), 185–206. https://doi.org/10.1111/zph.12549

Diblasi, A., & Bowman, A. (2001). On the use of variogram in checking for independence in spatial data. *Biometrics*, *March*, 211–218.

Diggle, P. J., & Ribeiro Jr, P. J. (2007). *Model-based geostatistics* (Springer S). Springer Science+Business Media, LLC 2007.

Disanto, G., Hall, C., Lucas, R., Ponsonby, A. L., Berlanga-Taylor, A. J., Giovannoni, G., & Ramagopalan, S. V. (2013). Assessing interactions between HLA-DRB1*15 and infectious mononucleosis on the risk of multiple sclerosis. *Multiple Sclerosis Journal*, *19*(10), 1355–1358. https://doi.org/10.1177/1352458513477231

Dogan, I., Onen, H. I., Yurdakul, A. S., Konac, E., Ozturk, C., Varol, A., & Ekmekci, A. (2009). Polymorphisms in the vitamin D receptor gene and risk of lung cancer. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*, *15*(8), BR232-42. http://www.ncbi.nlm.nih.gov/pubmed/19644412

Dolei, A., Ibba, G., Piu, C., & Serra, C. (2019). Expression of HERV genes as possible biomarker and target in neurodegenerative diseases. *International Journal of Molecular Sciences*, *20*(15). https://doi.org/10.3390/ijms20153706

Dunning, A. M., McBride, S., Gregory, J., Durocher, F., Foster, N. A., Healey, C. S., Smith, N., Pharoah, P. D. P., Luben, R. N., Easton, D. F., & Ponder, B. A. J. (1999). No association between androgen or vitamin D receptor gene polymorphisms and risk of breast cancer. *Carcinogenesis*, *20*(11), 2131–2135. https://doi.org/10.1093/carcin/20.11.2131

Ebers, G. C. (2008). Environmental factors and multiple sclerosis. *Lancet Neurology*, *7*(March), 268–277. https://doi.org/10.1586/14737175.2013.865866

Emmanouilidou, E., Galli-Tsinopoulou, A., Kyrgios, I., Gbandi, E., & Goulas, A. (2015). Common VDR polymorphisms and idiopathic short stature in children from northern Greece. *Hippokratia*, *19*(1), 25–29. http://www.ncbi.nlm.nih.gov/pubmed/26435642

Farinotti, M., Simi, S., C, D. P., Mcdowell, N., Brait, L., Lupo, D., & Filippini, G. (2020). Dietary interventions for multiple sclerosis (Review). *The Cochrane Library of Systematic Reviews*, *5*, 1–97. https://doi.org/10.1002/14651858.CD004192.pub4.www.cochranelibrary.com

Fernandez-Rozadilla, C., Alvarez-Barona, M., Schamschula, E., Bodo, S., Lopez-Novo, A., Dacal, A., Calviño-Costas, C., Lancho, A., Amigo, J., Bello, X., Cameselle-Teijeiro, J. M., Carracedo, A.,

Colas, C., Muleris, M., Wimmer, K., & Ruiz-Ponte, C. (2019). Early colorectal cancers provide new evidence for a lynch syndrome-to-CMMRD phenotypic continuum. *Cancers*, *11*(8), 1–11. https://doi.org/10.3390/cancers11081081

Fernandez-Rozadilla, C., Cazier, J. B., Tomlinson, I. P., Carvajal-Carmona, L. G., Palles, C., Lamas, M. J., Baiget, M., López-Fernández, L. A., Brea-Fernández, A., Abulí, A., Bujanda, L., Clofent, J., Gonzalez, D., Xicola, R., Andreu, M., Bessa, X., Jover, R., Llor, X., Moreno, V., … Ruiz-Ponte, C. (2013). A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-55

Flachenecker, P., Buckow, K., Pugliatti, M., Kes, V. B., Battaglia, M. a, Boyko, A., Ellenberger, D., Eskic, D., Ford, D., Friede, T., Fuge, J., Glaser, A., Hillert, J., Holloway, E., Ioannidou, E., & Kappos, L. (2014). Multiple sclerosis registries in Europe - results of a systematic survey. *Multiple Sclerosis (Houndmills, Basingstoke, England)*. https://doi.org/10.1177/1352458514528760

Fromont, A., Binquet, C., Sauleau, E. A., Fournel, I., Bellisario, A., Adnet, J., Weill, A., Vukusic, S., Confavreux, C., Debouverie, M., Clerc, L., Bonithon-Kopp, C., & Moreau, T. (2010). Geographic variations of multiple sclerosis in France. *Brain*, *133*(7), 1889–1899. https://doi.org/10.1093/brain/awq134

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). Minimac2: Faster genotype imputation. *Bioinformatics*, *31*(5), 782–784. https://doi.org/10.1093/bioinformatics/btu704

Gapska, P., Scott, R. J., Serrano-Fernandez, P., Huzarski, T., Byrski, T., Kładny, J., Gronwald, J., Górski, B., Cybulski, C., Lubinski, J., & Dębniak, T. (2009). Vitamin D receptor variants and breast cancer risk in the Polish population. *Breast Cancer Research and Treatment*, *115*(3), 629–633. https://doi.org/10.1007/s10549-008-

0107-1

Gapska, P., Scott, R. J., Serrano-Fernandez, P., Mirecka, A., Rassoud, I., Górski, B., Cybulski, C., Huzarski, T., Byrski, T., Nagay, L., Maleszka, R., Sulikowski, M., Lubinski, J., & Debniak, T. (2009). Vitamin D receptor variants and the malignant melanoma risk: A population-based study. *Cancer Epidemiology*, *33*(2), 103–107. https://doi.org/10.1016/j.canep.2009.06.006

García-Martín, E., Agúndez, J. A. G., Martínez, C., Benito-León, J., Millán-Pascual, J., Calleja, P., Díaz-Sánchez, M., Pisa, D., Turpín-Fenoll, L., Alonso-Navarro, H., Ayuso-Peralta, L., Torrecillas, D., Plaza-Nieto, J. F., & Jiménez-Jiménez, F. J. (2013). Vitamin D3 Receptor (VDR) Gene rs2228570 (Fok1) and rs731236 (Taq1) Variants Are Not Associated with the Risk for Multiple Sclerosis: Results of a New Study and a Meta-Analysis. *PLoS ONE*, *8*(6), e65487. https://doi.org/10.1371/journal.pone.0065487

Geginat, J., Paroni, M., Pagani, M., Galimberti, D., De Francesco, R., Scarpini, E., & Abrignani, S. (2017). The Enigmatic Role of Viruses in Multiple Sclerosis: Molecular Mimicry or Disturbed Immune Surveillance? *Trends in Immunology*, *38*(7), 498–512. https://doi.org/10.1016/j.it.2017.04.006

Ghattaoraya, G. S., Dundar, Y., González-Galarza, F. F., Maia, M. H. T., Santos, E. J. M., Da Silva, A. L. S., McCabe, A., Middleton, D., Alfirevic, A., Dickson, R., & Jones, A. R. (2016). A web resource for mining HLA associations with adverse drug reactions: HLA-ADR. *Database*, *2016*, 1–10. https://doi.org/10.1093/database/baw069

Gianfrancesco, M. A., Glymour, M. M., Walter, S., Rhead, B., Shao, X., Shen, L., Quach, H., Hubbard, A., Jónsdóttir, I., Stefánsson, K., Strid, P., Hillert, J., Hedström, A., Olsson, T., Kockum, I., Schaefer, C., Alfredsson, L., & Barcellos, L. F. (2017). Causal Effect of Genetic Variants Associated with Body Mass Index on Multiple Sclerosis Susceptibility. *American Journal of*

Epidemiology, *185*(3), 162–171. https://doi.org/10.1093/aje/kww120

Gianfrancesco, M., Acuna, B., Shen, L., Briggs, F., Quach, H., Bellesis, K., Bernstein, A., Hedstrom, A., Kockum, I., Alfredsson, L., Olsson, T., Schaefer, C., & Barcellos, L. (2014). Obesity during childhood and adolescence increases susceptibility to multiple sclerosis after accounting for established genetic and environmental risk factors. *Obes Res Clin Pract.*, *8*(5), 1–7. https://doi.org/10.1038/jid.2014.371

Giorgi, E., Diggle, P. J., Snow, R. W., & Noor, A. M. (2018). Geostatistical Methods for Disease Mapping and Visualisation Using Data from Spatio-temporally Referenced Prevalence Surveys. *International Statistical Review*, *86*(3), 571–597. https://doi.org/10.1111/insr.12268

Golden, L. C., & Voskuhl, R. (2017). The importance of studying sex differences in disease: The example of multiple sclerosis. *Journal of Neuroscience Research*, *95*(1–2), 633–643. https://doi.org/10.1002/jnr.23955

González-Galarza, F. F., Takeshita, L. Y. C., Santos, E. J. M., Kempson, F., Maia, M. H. T., Da Silva, A. L. S., Teles E Silva, A. L., Ghattaoraya, G. S., Alfirevic, A., Jones, A. R., & Middleton, D. (2015). Allele frequency net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Research*, *43*(D1), D784–D788. https://doi.org/10.1093/nar/gku1166

Goodman, M., Naiman, J. S., Goodman, D., & LaKind, J. S. (2012). Cancer clusters in the USA: What do the last twenty years of state and federal investigations tell us? *Critical Reviews in Toxicology*, *42*(6), 474–490. https://doi.org/10.3109/10408444.2012.675315

Grant, W. B. (2010). An ecological study of cancer incidence and mortality rates in France with respect to latitude, an index for

vitamin D production. *Dermato-Endocrinology*, *2*(2), 62–67. https://doi.org/10.4161/derm.2.2.13624

Grant, W. B., & Mascitelli, L. (2012). Evidence that the north-south gradient of multiple sclerosis may not have disappeared. *Journal of the Neurological Sciences*, *315*(1–2), 178–179. https://doi.org/10.1016/j.jns.2012.01.002

Habes, M., Grothe, M. J., Tunc, B., McMillan, C., Wolk, D. A., & Davatzikos, C. (2020). Disentangling Heterogeneity in Alzheimer's Disease and Related Dementias Using Data-Driven Methods. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2020.01.016

Haghighi, S., Lekman, A., Nilsson, S., Blomqvist, M., & Andersen, O. (2013). Increased CSF sulfatide levels and serum glycosphingolipid antibody levels in healthy siblings of multiple sclerosis patients. *Journal of the Neurological Sciences*, *326*(1–2), 35–39. https://doi.org/10.1016/j.jns.2013.01.007

Hammond, S. R., English, D. R., & McLeod, J. G. (2000). The age-range of risk of developing multiple sclerosis: evidence from a migrant population in Australia. *Brain : A Journal of Neurology*, *123 (Pt 5*(2000), 968–974. https://doi.org/10.1093/brain/123.5.968

Hancock, A. M., Witonsky, D. B., Gordon, A. S., Eshel, G., Pritchard, J. K., Coop, G., & Di Rienzo, A. (2008). Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, *4*(2). https://doi.org/10.1371/journal.pgen.0040032

Harris, V. K., & Sadiq, S. A. (2014). Biomarkers of Therapeutic Response in Multiple Sclerosis: Current Status. *Molecular Diagnosis & Therapy*, *18*(6), 605–617. https://doi.org/10.1007/s40291-014-0117-0

Hart, P. H., & Gorman, S. (2013). Exposure to UV wavelengths in

sunlight suppresses immunity. To what extent is UV-induced vitamin D3 the mediator responsible? *Clinical Biochemist Reviews*, *34*(February), 3–13.

Hedström, A. K., Olsson, T., & Alfredsson, L. (2012). High body mass index before age 20 is associated with increased risk for multiple sclerosis in both men and women. *Multiple Sclerosis Journal*, *18*(9), 1334–1336. https://doi.org/10.1177/1352458512436596

Hedström, A., Katsoulis, M., Hössjer, O., Bomfim, I. L., Oturai, A., Sondergaard, H. B., Sellebjerg, F., Ullum, H., Thørner, L. W., Gustavsen, M. W., Harbo, H. F., Obradovic, D., Gianfrancesco, M. A., Barcellos, L. F., Schaefer, C. A., Hillert, J., Kockum, I., Olsson, T., & Alfredsson, L. (2017). The interaction between smoking and HLA genes in multiple sclerosis: replication and refinement. *European Journal of Epidemiology*, 1–11. https://doi.org/10.1007/s10654-017-0250-2

Heine, G., Hoefer, N., Franke, A., Nöthling, U., Schumann, R. R., Hamann, L., & Worm, M. (2013). Association of vitamin D receptor gene polymorphisms with severe atopic dermatitis in adults. *British Journal of Dermatology*, *168*(4), 855–858. https://doi.org/10.1111/bjd.12077

Hewes, D., Tatomir, A., Kruszewski, A. M., Rao, G., Tegla, C. A., Ciriello, J., Nguyen, V., Royal, W., Bever, C., Rus, V., & Rus, H. (2017). SIRT1 as a potential biomarker of response to treatment with glatiramer acetate in multiple sclerosis. *Experimental and Molecular Pathology*, *102*(2), 191–197. https://doi.org/10.1016/j.yexmp.2017.01.014

Horst-Sikorska, W., Ignaszak-Szczepaniak, M., Marcinkowska, M., Kaczmarek, M., Stajgis, M., & Slomski, R. (2008). Association analysis of vitamin D receptor gene polymorphisms with bone mineral density in young women with Graves' disease. *Acta Biochimica Polonica*, *55*(2), 371–380. https://doi.org/10.18388/abp.2008_3085

Houzen, H., Kondo, K., Horiuchi, K., & Niino, M. (2018). Consistent increase in the prevalence and female ratio of multiple sclerosis over 15 years in northern Japan. *European Journal of Neurology*, *25*(2), 334–339. https://doi.org/10.1111/ene.13506

Howell, A. E., Robinson, J. W., Wootton, R. E., Mcaleenan, A., Tsavachidis, S., Ostrom, Q. T., Bondy, M., Armstrong, G., Relton, C., Haycock, P., Martin, R. M., Zheng, J., & Kurian, K. M. (2020). Testing for causality between systematically identified risk factors and glioma : a Mendelian randomization study. *BMC Cancerncer*, *20*, 1–11.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, *1*(6), 457–470. https://doi.org/10.1534/g3.111.001198

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6). https://doi.org/10.1371/journal.pgen.1000529

Http://re.jrc.ec.europa.eu/pvgis/apps4/pvest.php. (2016). *http://re.jrc.ec.europa.eu/pvgis/apps4/pvest.php*.

http://www.allelefrequencies.net. (n.d.). *http://www.allelefrequencies.net*.

Hughes, D. J., Hlavatá, I., Soucek, P., Pardini, B., Naccarati, A., Vodickova, L., Jenab, M., & Vodicka, P. (2011). Variation in the vitamin D receptor gene is not associated with risk of colorectal cancer in the Czech Republic. *Journal of Gastrointestinal Cancer*, *42*(3), 149–154. https://doi.org/10.1007/s12029-010-9168-6

Huld, T., Müller, R., & Gambardella, A. (2012). A new solar radiation database for estimating PV performance in Europe and Africa. *Solar Energy*, *86*(6), 1803–1815. https://doi.org/10.1016/j.solener.2012.03.006

Irizar, H., Muñoz-Culla, M., Zuriarrain, O., Goyenechea, E., Castillo-Triviño, T., Prada, A., Saenz-Cuesta, M., De Juan, D., Lopez De Munain, A., Olascoaga, J., & Otaegui, D. (2012). HLA-DRB1*15:01 and multiple sclerosis: A female association? *Multiple Sclerosis Journal*, *18*(5), 569–577. https://doi.org/10.1177/1352458511426813

Isobe, N., Damotte, V., Lo Re, V., Ban, M., Pappas, D., Guillot-Noel, L., Rebeix, I., Compston, A., Mack, T., Cozen, W., Fontaine, B., Hauser, S. L., Oksenberg, J. R., Sawcer, S., & Gourraud, P. A. (2013). Genetic burden in multiple sclerosis families. *Genes and Immunity*, *14*(7), 434–440. https://doi.org/10.1038/gene.2013.37

Jablonski, N. G., & Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution*, *39*, 57–106. https://doi.org/10.1006/jhev.2000.0403

Jablonski, N. G., & Chaplin, G. (2012). Human skin pigmentation, migration and disease susceptibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(February), 785–792. https://doi.org/10.1098/rstb.2011.0308

Jombart, T., Eggo, R. M., Dodd, P. J., & Balloux, F. (2011). Reconstructing disease outbreaks from genetic data: A graph approach. *Heredity*, *106*(2), 383–390. https://doi.org/10.1038/hdy.2010.78

Källberg, H., Vieira, V., Holmqvist, M., Hart, J., Costenbader, K., Bengtsson, C., Klareskog, L., Karlson, E., & Alfredsson, L. (2013). Regional differences regarding risk of developing Rheumatoid Arthritis in Stockholm County, Sweden: Results from the Swedish Epidemiological Investigation of Rheumatoid

Arthritis (EIRA) study. *Scand J Rheumatol*, *42*(5). https://doi.org/10.1038/jid.2014.371

Karami, S., Brennan, P., Hung, R. J., Boffetta, P., Toro, J., Wilson, R. T., Zaridze, D., Navratilova, M., Chatterjee, N., Mates, D., Janout, V., Kollarova, H., Bencko, V., Szeszenia-Dabrowska, N., Holcatova, I., Moukeria, A., Welch, R., Chanock, S., Rothman, N., … Moore, L. E. (2008). Vitamin D receptor polymorphisms and renal cancer risk in Central and Eastern Europe. *Journal of Toxicology and Environmental Health - Part A: Current Issues*, *71*(6), 367–372. https://doi.org/10.1080/15287390701798685

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses* (version 1.0.7). https://cran.r-project.org/package=factoextra%0A

Kempiska-Podhorecka, A., Wunsch, E., Jarowicz, T., Raszeja-Wyszomirska, J., Loniewska, B., Kaczmarczyk, M., Milkiewicz, M., & Milkiewicz, P. (2012). Vitamin D receptor polymorphisms predispose to primary biliary cirrhosis and severity of the disease in polish population. *Gastroenterology Research and Practice*, *2012*. https://doi.org/10.1155/2012/408723

Kingwell, E., Marriott, J. J., Jetté, N., Pringsheim, T., Makhani, N., Morrow, S. a, Fisk, J. D., Evans, C., Béland, S. G., Kulaga, S., Dykeman, J., Wolfson, C., Koch, M. W., & Marrie, R. A. (2013). Incidence and prevalence of multiple sclerosis in Europe: a systematic review. *BMC Neurology*, *13*, 128. https://doi.org/10.1186/1471-2377-13-128

Koch-Henriksen, N., & Sorensen, P. S. (2011). Why does the north-south gradient of incidence of multiple sclerosis seem to have disappeared on the Northern hemisphere? *Journal of the Neurological Sciences*, *311*(1–2), 58–63. https://doi.org/10.1016/j.jns.2011.09.003

Kragt, J., van Amerongen, B., Killestein, J., Dijkstra, C., Uitdehaag, B.,

Polman, C., & Lips, P. (2009). Higher levels of 25-hydroxyvitamin D are associated with a lower incidence of multiple sclerosis only in women. *Multiple Sclerosis (Houndmills, Basingstoke, England)*, *15*(March 2008), 9–15. https://doi.org/10.1177/1352458508095920

Kujundzic, B., Zeljic, K., Supic, G., Magic, M., Stanimirovic, D., Ilic, V., Jovanovic, B., & Magic, Z. (2016). Association of vdr, cyp27b1, cyp24a1 and mthfr gene polymorphisms with oral lichen planus risk. *Clinical Oral Investigations*, *20*(4), 781–789. https://doi.org/10.1007/s00784-015-1572-7

Kumar, A., Cocco, E., Atzori, L., Marrosu, M. G., & Pieroni, E. (2013). Structural and Dynamical Insights on HLA-DR2 Complexes That Confer Susceptibility to Multiple Sclerosis in Sardinia: A Molecular Dynamics Simulation Study. *PLoS ONE*, *8*(3), 1–13. https://doi.org/10.1371/journal.pone.0059711

Kurt, O., Yilmaz-Aydogan, H., Uyar, M., Isbir, T., Seyhan, M. F., & Can, A. (2012). Evaluation of ERα and VDR gene polymorphisms in relation to bone mineral density in Turkish postmenopausal women. *Molecular Biology Reports*, *39*(6), 6723–6730. https://doi.org/10.1007/s11033-012-1496-0

Kurtzke, J. F. (2013). Epidemiology in multiple sclerosis: a pilgrim's progress. *Brain*, *136*(9), 2904–2917. https://doi.org/10.1093/brain/awt220

Łaczmański, Ł., Jakubik, M., Bednarek-Tupikowska, G., Rymaszewska, J., Słoka, N., & Lwow, F. (2015). Vitamin D receptor gene polymorphisms in Alzheimer's disease patients. *Experimental Gerontology*, *69*, 142–147. https://doi.org/10.1016/j.exger.2015.06.012

Langer-Gould, A., Lucas, R., Xiang, A. H., Chen, L. H., Wu, J., Gonzalez, E., Haraszti, S., Smith, J. B., Quach, H., & Barcellos, L. F. (2018). MS sunshine study: Sun exposure but not vitamin D is

associated with multiple sclerosis risk in blacks and hispanics. *Nutrients*, *10*(3), 1–14. https://doi.org/10.3390/nu10030268

Lema Casal, A. (2015). *Comparación de técnicas de estimación de datos genéticos: imputación genética vs interpolación espacial*.

Lemke, L. D., Lamerato, L. E., Xu, X., Booza, J. C., Reiners, J. J., Raymond Iii, D. M., Villeneuve, P. J., Lavigne, E., Larkin, D., & Krouse, H. J. (2014). Geospatial relationships of air pollution and acute asthma events across the Detroit-Windsor international border: Study design and preliminary results. *Journal of Exposure Science and Environmental Epidemiology*, *24*(4), 346–357. https://doi.org/10.1038/jes.2013.78

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834. https://doi.org/10.1002/gepi.20533

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype Imputation. *Annual Review of Genomics and Human Genetics*, *10*(1), 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

López-Cortés, A., Cabrera-Andrade, A., Vázquez-Naya, J. M., Pazos, A., Gonzáles-Díaz, H., Paz-y-Miño, C., Guerrero, S., Pérez-Castillo, Y., Tejera, E., & Munteanu, C. R. (2019). Prediction of breast cancer proteins using molecular descriptors and artificial neural networks: a focus on cancer immunotherapy proteins, metastasis driver proteins, and RNA-binding proteins. *BioRxiv*, 840108. https://doi.org/10.1101/840108

Lublin, Fred D;, & Reingold, S. C. (1996). *Defining the clinical course of multiple sclerosis 1996.pdf*. 907–912.

Lublin, Fred D, Reingold, S. C., Cohen, J. a, Cutter, G. R., Thompson, A. J., Wolinsky, J. S., Fox, R. J., Freedman, M. S., Goodman, A.

D., & Lubetzki, C. (2014). *Defining the clinical course of multiple sclerosis : The 2013 revisions Defining the clinical course of multiple sclerosis The 2013 revisions*. 1–10. https://doi.org/10.1212/WNL.0000000000000560

Lundin, A. C., Söderkvist, P., Eriksson, B., Bergman-Jungeström, M., & Wingren, S. (1999). Association of breast cancer progression with a vitamin D receptor gene polymorphism. *Cancer Research*, *59*(10), 2332–2334.

Maalej, A., Petit-Teixeira, E., Chabchoub, G., Hamad, M. Ben, Rebai, A., Farid, N. R., Cornelis, F., & Ayadi, H. (2008). Lack of association of VDR gene polymorphisms with thyroid autoimmune disorders: Familial and case/control studies. *Journal of Clinical Immunology*, *28*(1), 21–25. https://doi.org/10.1007/s10875-007-9124-9

Madsen, L., Ruppert, D., & Altman, N. (2008). Regression with spatially misaligned data. *Environmetrics*, *19*, 453–467. https://doi.org/10.1002/env

Mameli, G., Poddighe, L., Mei, A., Uleri, E., Sotgiu, S., Serra, C., Manetti, R., & Dolei, A. (2012). Expression and Activation by Epstein Barr Virus of Human Endogenous Retroviruses-W in Blood Cells and Astrocytes: Inference for Multiple Sclerosis. *PLoS ONE*, *7*(9), 1–13. https://doi.org/10.1371/journal.pone.0044991

Many, N., Stickel, F., Schmitt, J., Stieger, B., Soyka, M., Frei, P., Götze, O., Müllhaupt, B., & Geier, A. (2012). Genetic variations in bile acid homeostasis are not overrepresented in alcoholic cirrhosis compared to patients with heavy alcohol abuse and absent liver disease. *Mutagenesis*, *27*(5), 567–572. https://doi.org/10.1093/mutage/ges020

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499–

511. https://doi.org/10.1038/nrg2796

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*(7), 906–913. https://doi.org/10.1038/ng2088

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), 1–10. https://doi.org/10.1002/mpr.1608

Marrosu, M. G., Murru, R., Murru, M. R., Costa, G., Zavattari, P., Whalen, M., Cocco, E., Mancosu, C., Schirru, L., Solla, E., Fadda, E., Melis, C., Porru, I., Rolesu, M., & Cucca, F. (2001). Dissection of the HLA association with multiple sclerosis in the founder isolated population of Sardinia. *Human Molecular Genetics*, *10*(25), 2907–2916.

Marrosu, M., Murru, M., Costa, G., Cucca, F., Sotgiu, S., Rosati, G., & Muntoni, F. (1997). Multiple sclerosis in Sardinia is associated and in linkage disequilibrium with HLA-DR3 and -DR4 alleles. *Am. J. Hum. Genet.*, *61*, 454–457.

Martin, R. J. L., McKnight, A. J., Patterson, C. C., Sadlier, D. M., Maxwell, A. P., & Warren 3/UK GoKinD Study Group. (2010). A rare haplotype of the vitamin D receptor gene is protective against diabetic nephropathy. *Nephrology, Dialysis, Transplantation : Official Publication of the European Dialysis and Transplant Association - European Renal Association*, *25*(2), 497–503. https://doi.org/10.1093/ndt/gfp515

McAllister, K., Mechanic, L. E., Amos, C., Aschard, H., Blair, I. A., Chatterjee, N., Conti, D., Gauderman, W. J., Hsu, L., Hutter, C. M., Jankowska, M. M., Kerr, J., Kraft, P., Montgomery, S. B., Mukherjee, B., Papanicolaou, G. J., Patel, C. J., Ritchie, M. D.,

Ritz, B. R., … Witte, J. S. (2017). Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *American Journal of Epidemiology*, *186*(7), 753–761. https://doi.org/10.1093/aje/kwx227

Mcdonald, W., Compston, A., Edan, G., Goodkin, D., HP, H., Lublin, F., & Mcfarland, H. (2001). Recommended diagnostic criteria for multiple sclerosis : guidelines from the International Panel on the diagnosis of multiple sclerosis. *Annals of Neurology*, *59*(April), 11456302. https://doi.org/10.1002/ana.1032

McKinney, B. A., Reif, D. M., Ritchie, M. D., & Moore, J. H. (2006). Machine learning for detecting gene-gene interactions: A review. *Applied Bioinformatics*, *5*(2), 77–88. https://doi.org/10.2165/00822942-200605020-00002

Milosevic, E., Dujmovic, I., Markovic, M., Mesaros, S., Rakocevic, G., Drulovic, J., Stojkovic, M. M., & Popadic, D. (2015). Higher expression of IL-12Rβ2 is associated with lower risk of relapse in relapsing-remitting multiple sclerosis patients on interferon-β1b therapy during 3-year follow-up. *Journal of Neuroimmunology*, *287*, 64–70. https://doi.org/10.1016/j.jneuroim.2015.07.011

Mirza, A., & Mao-Draayer, Y. (2016). The gut microbiome and microbial translocation in multiple sclerosis. *Clinical Immunology*. https://doi.org/10.1016/j.clim.2017.03.001

Mische, L. J., & Mowry, E. M. (2018). The Evidence for Dietary Interventions and Nutritional Supplements as Treatment Options in Multiple Sclerosis: a Review. *Current Treatment Options in Neurology*, *20*(4). https://doi.org/10.1007/s11940-018-0494-5

Mokry, L. E., Ross, S., Timpson, N. J., Sawcer, S., Davey Smith, G., & Richards, J. B. (2016). Obesity and Multiple Sclerosis: A Mendelian Randomization Study. *PLoS Medicine*, *13*(6), 1–16. https://doi.org/10.1371/journal.pmed.1002053

Morandi, E., Tanasescu, R., Tarlinton, R. E., Constantinescu, C. S., Zhang, W., Tench, C., & Gran, B. (2017). *The association between human endogenous retroviruses and multiple sclerosis : A systematic review and meta-analysis*. *66*, 1–18. https://doi.org/10.1371/journal.pone.0172415

Mostowska, A., Lianeri, M., Wudarski, M., Olesińska, M., & Jagodziński, P. P. (2013). Vitamin D receptor gene BsmI, FokI, ApaI and TaqI polymorphisms and the risk of systemic lupus erythematosus. *Molecular Biology Reports*, *40*(2), 803–810. https://doi.org/10.1007/s11033-012-2118-6

Moutsianas, L., Jostins, L., Beecham, A. H., Dilthey, A. T., Xifara, D. K., Ban, M., Shah, T. S., Patsopoulos, N. A., Alfredsson, L., Anderson, C. A., Attfield, K. E., Baranzini, S. E., Barrett, J., Binder, T. M. C., Booth, D., Buck, D., Celius, E. G., Cotsapas, C., D'Alfonso, S., … McVean, G. (2015). Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nature Genetics*, *47*(10), 1107–1113. https://doi.org/10.1038/ng.3395

MSIF and WHO. (2008). Atlas multiple sclerosis resources in the world 2008. In *ISBN 978 92 4 156375 8*.

MSIF and WHO. (2013). *Atlas of MS 2013* (Multiple Sclerosis International Federation 2013 (ed.)).

Munger, K. L., Chitnis, T., Ascherio, A., & Munger, K. L. (2013). *Body size and risk of MS in two cohorts of US women*. https://doi.org/10.1212/WNL.0b013e3181c0d6e0

Munger, K. L., Levin, L. I., Hollis, B. W., Howard, N. S., & Ascherio, A. (2006). Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. *JAMA : The Journal of the American Medical Association*, *296*(23), 2832–2838. https://doi.org/10.1001/jama.296.23.2832

Novembre, J., & Di Rienzo, A. (2009). Spatial patterns of variation due

to natural selection in humans. *Nature Reviews Genetics*, *10*(11), 745–755. https://doi.org/10.1038/nrg2632

O'Gorman, C., Lin, R., Stankovich, J., & Broadley, S. a. (2012). Modelling genetic susceptibility to multiple sclerosis with family data. *Neuroepidemiology*, *40*, 1–12. https://doi.org/10.1159/000341902

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., & Simianer, H. (2011). Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics*, *188*(3), 695–708. https://doi.org/10.1534/genetics.111.128694

Olitsky, P. K., & Yager, R. H. (1949). *Experimental disseminated encephalomyelitits in white mice*. *6*.

Ozonoff, A., Webster, T., Vieira, V., Weinberg, J., Ozonoff, D., & Aschengrau, A. (2005). Cluster detection methods applied to the Upper Cape Cod cancer data. *Environmental Health: A Global Access Science Source*, *4*, 1–9. https://doi.org/10.1186/1476-069X-4-19

Panierakis, C., Goulielmos, G., Mamoulakis, D., Petraki, E., Papavasiliou, E., & Galanakis, E. (2009). Vitamin D receptor gene polymorphisms and susceptibility to type 1 diabetes in Crete, Greece. *Clinical Immunology*, *133*(2), 276–281. https://doi.org/10.1016/j.clim.2009.08.004

Patsopoulos, N. a., Barcellos, L. F., Hintzen, R. Q., Schaefer, C., van Duijn, C. M., Noble, J. a., Raj, T., Gourraud, P. A., Stranger, B. E., Oksenberg, J., Olsson, T., Taylor, B. V., Sawcer, S., Hafler, D. a., Carrington, M., De Jager, P. L., & de Bakker, P. I. W. (2013). Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. *PLoS Genetics*, *9*(11). https://doi.org/10.1371/journal.pgen.1003926

Pierrot-Deseilligny, C., & Souberbielle, J.-C. (2010). Is hypovitaminosis D one of the environmental risk factors for multiple sclerosis? *Brain : A Journal of Neurology*, *133*, 1869–1888. https://doi.org/10.1093/brain/awq147

Pierrot-Deseilligny, C., & Souberbielle, J.-C. (2013). Contribution of vitamin D insufficiency to the pathogenesis of multiple sclerosis. *Therapeutic Advances in Neurological Disorders*, *6*(Paris V), 81–116. https://doi.org/10.1177/1756285612473513

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F. D., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A. J., Waubant, E., Weinshenker, B., & Wolinsky, J. S. (2011). Diagnostic criteria for multiple sclerosis: 2010 Revisions to the McDonald criteria. *Annals of Neurology*, *69*(2), 292–302. https://doi.org/10.1002/ana.22366

Polman, C., Reingold, S., Gilles, E., Filippi, M., Hartung, H., Kappos, L., Lublin, F., Metz, L., Mcfarland, H., O'Connor, P., Sandberg-Wollheim, M., Thompson, A., Weinshenker, B., & Wolinsky, J. (2005). Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria." *Ann Neurol*, *58*(1), 840–846.

Poser, C. M., Paty, D. W., Scheinberg, L., McDonald, W. I., Davis, F. A., Ebers, G. C., Johnson, K. P., Sibley, W. A., Silberberg, D. H., & Tourtellotte, W. W. (1983). New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Annals of Neurology*, *13*(3), 227–231. https://doi.org/10.1002/ana.410130302

Pugliatti, M., Rosati, G., Carton, H., Riise, T., Drulovic, J., Vécsei, L., & Milanov, I. (2006). The epidemiology of multiple sclerosis in Europe. *European Journal of Neurology*, *13*, 700–722. https://doi.org/10.1111/j.1468-1331.2006.01342.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J.,

& Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R Foundation for Statistical Computing. (2018). *R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Rajeevan, H., Osier, M. V., Cheung, K. H., Deng, H., Druskin, L., Heinzen, R., Kidd, J. R., Stein, S., Pakstis, A. J., Tosches, N. P., Yeh, C. C., Miller, P. L., & Kidd, K. K. (2003). ALFRED: The ALelle FREquency Database. Update. *Nucleic Acids Research*, *31*(1), 270–271. https://doi.org/10.1093/nar/gkg043

Rajeevan, Haseena, Cheung, K.-H., Gadagkar, R., Stein, S., Soundararajan, U., Kidd, J. R., Pakstis, A. J., Miller, P. L., & Kidd, K. K. (2005). ALFRED: An Allele Frequency Database for Microevolutionary Studies. *Evolutionary Bioinformatics*, *1*, 117693430500100. https://doi.org/10.1177/117693430500100006

Rajeevan, Haseena, Soundararajan, U., Kidd, J. R., Pakstis, A. J., & Kidd, K. K. (2012). ALFRED: An allele frequency resource for research and teaching. *Nucleic Acids Research*, *40*(D1), 1010–1015. https://doi.org/10.1093/nar/gkr924

Ramagopalan, S. V., Dobson, R., Meier, U. C., & Giovannoni, G. (2010). Multiple sclerosis: risk factors, prodromes, and potential causal pathways. *The Lancet Neurology*, *9*(7), 727–739. https://doi.org/10.1016/S1474-4422(10)70094-6

Ramagopalan, S. V., Maugeri, N. J., Handunnetthi, L., Lincoln, M. R., Orton, S. M., Dyment, D. a., DeLuca, G. C., Herrera, B. M., Chao, M. J., Sadovnick, a. D., Ebers, G. C., & Knight, J. C. (2009). Expression of the multiple sclerosis-associated MHC class II allele

HLA-DRB1*1501 is regulated by vitamin D. *PLoS Genetics*, *5*(2), 1–6. https://doi.org/10.1371/journal.pgen.1000369

Ramos-Lopez, E., Kurylowicz, A., Bednarczuk, T., Paunkovic, J., Seidl, C., & Badenhoop, K. (2005). Vitamin D receptor polymorphisms are associated with Graves' disease in German and Polish but not in Serbian patients. *Thyroid*, *15*(10), 1125–1130. https://doi.org/10.1089/thy.2005.15.1125

Randerson-Moor, J. A., Taylor, J. C., Elliott, F., Chang, Y. M., Beswick, S., Kukalizch, K., Affleck, P., Leake, S., Haynes, S., Karpavicius, B., Marsden, J., Gerry, E., Bale, L., Bertram, C., Field, H., Barth, J. H., Silva, I. dos S., Swerdlow, A., Kanetsky, P. A., … Bishop, J. A. N. (2009). Vitamin D receptor gene polymorphisms, serum 25-hydroxyvitamin D levels, and melanoma: UK case-control comparisons and a meta-analysis of published VDR data. *European Journal of Cancer*, *45*(18), 3271–3281. https://doi.org/10.1016/j.ejca.2009.06.011

Ribeiro Jr, P. J., & Diggle, P. J. (2018). *geoR: Analysis of Geostatistical Data*.

Riccio, P., & Rossano, R. (2017). Diet, Gut Microbiota, and Vitamins D + A in Multiple Sclerosis. *Neurotherapeutics*, *1*, 1–17. https://doi.org/http://dx.doi.org/10.1007/s13311-017-0581-4

Risco, J., Maldonado, H., Luna, L., Osada, J., Ruiz, P., Juarez, A., & Vizcarra, D. (2011). Latitudinal prevalence gradient of multiple sclerosis in Latin America. *Multiple Sclerosis Journal*, *17*(9), 1055–1059. https://doi.org/10.1177/1352458511405562

Rucevic, I., Stefanic, M., Tokic, S., Vuksic, M., Glavas-Obrovac, L., & Barisic-Drusko, V. (2012). Lack of association of vitamin D receptor gene 3′-haplotypes with psoriasis in Croatian patients. *Journal of Dermatology*, *39*(1), 58–62. https://doi.org/10.1111/j.1346-8138.2011.01296.x

Sackton, T. B., & Hartl, D. L. (2016). Genotypic Context and Epistasis in Individuals and Populations. *Cell*, *166*(2), 279–287. https://doi.org/10.1016/j.cell.2016.06.047

Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C. a, Patsopoulos, N. a, Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S. E., Edkins, S., Gray, E., Booth, D. R., Potter, S. C., Goris, A., Band, G., Oturai, A. B., Strange, A., Saarela, J., … Compston, A. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, *476*(527), 214–219. https://doi.org/10.1038/nature10251

Schabenberguer, O., & Gotway, C. A. (2005). *Statistical methods for spatial data analysis* (B. P. Carlin, C. Chatfield, M. Tanner, & J. Zidek (eds.)). Champan & Hall/CRC.

Schootman, M., Nelson, E. J., Werner, K., Shacham, E., Elliott, M., Ratnapradipa, K., Lian, M., & McVay, A. (2016). Emerging technologies to measure neighborhood conditions in public health: Implications for interventions and next steps. *International Journal of Health Geographics*, *15*(1), 1–9. https://doi.org/10.1186/s12942-016-0050-z

Shi, H., Tan, S. jie, Zhong, H., Hu, W., Levine, A., Xiao, C. jie, Peng, Y., Qi, X. bin, Shou, W. hua, Ma, R. lin Z., Li, Y., Su, B., & Lu, X. (2009). Winter Temperature and UV Are Tightly Linked to Genetic Changes in the p53 Tumor Suppressor Pathway in Eastern Asia. *American Journal of Human Genetics*, *84*(4), 534–541. https://doi.org/10.1016/j.ajhg.2009.03.009

Signoriello, E., Lanzillo, R., Brescia Morra, V., Di Iorio, G., Fratta, M., Carotenuto, A., & Lus, G. (2016). Lymphocytosis as a response biomarker of natalizumab therapeutic efficacy in multiple sclerosis. *Multiple Sclerosis Journal*, *22*(7), 921–925. https://doi.org/10.1177/1352458515604381

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of

Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, *6*(5). https://doi.org/10.1371/journal.pone.0019467

Sillanpää, P., Hirvonen, A., Kataja, V., Eskelinen, M., Kosma, V. M., Uusitupa, M., Vainio, H., & Mitrunen, K. (2004). Vitamin D receptor gene polymorphism as an important modifier of positive family history related breast cancer risk. *Pharmacogenetics*, *14*(4), 239–245. https://doi.org/10.1097/00008571-200404000-00003

Simon, K. C., Munger, K. L., Kraft, P., Hunter, D. J., De Jager, P. L., & Ascherio, a. (2011). Genetic predictors of 25-hydroxyvitamin D levels and risk of multiple sclerosis. *Journal of Neurology*, *258*, 1676–1682. https://doi.org/10.1007/s00415-011-6001-5

Simpson, S., Blizzard, L., Otahal, P., Van der Mei, I., & Taylor, B. (2011). Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis. *Journal of Neurology, Neurosurgery, and Psychiatry*, *82*, 1132–1141. https://doi.org/10.1136/jnnp.2011.240432

Simpson, S., Taylor, B., Blizzard, L., Ponsonby, A. L., Pittas, F., Tremlett, H., Dwyer, T., Gies, P., & Van Der Mei, I. (2010). Higher 25-hydroxyvitamin D is associated with lower relapse risk in multiple sclerosis. *Annals of Neurology*, *68*(2), 193–203. https://doi.org/10.1002/ana.22043

Simpson, S., Wang, W., Otahal, P., Blizzard, L., Van Der Mei, I. A. F., & Taylor, B. V. (2019). Latitude continues to be significantly associated with the prevalence of multiple sclerosis: An updated meta-analysis. *Journal of Neurology, Neurosurgery and Psychiatry*, *90*(11), 1193–1200. https://doi.org/10.1136/jnnp-2018-320189

Sioka, C., Papakonstantinou, S., Markoula, S., Gkartziou, F., Georgiou, A., Georgiou, I., Pelidou, S. H., Kyritsis, A. P., & Fotopoulos, A. (2011). Vitamin D receptor gene polymorphisms in multiple

sclerosis patients in northwest Greece. *Journal of Negative Results in BioMedicine*, *10*(1). https://doi.org/10.1186/1477-5751-10-3

Sloan, C. D., Duell, E. J., Shi, X., Irwin, R., Andrew, A. S., Williams, S. M., & Moore, J. H. (2009). Ecogeographic genetic epidemiology. *Genetic Epidemiology*, *33*, 281–289. https://doi.org/10.1002/gepi.20386

Smedby, K. E., Eloranta, S., Duvefelt, K., Melbye, M., Humphreys, K., Hjalgrim, H., & Chang, E. T. (2011). Vitamin D receptor genotypes, ultraviolet radiation exposure, and risk of non-Hodgkin lymphoma. *American Journal of Epidemiology*, *173*(1), 48–54. https://doi.org/10.1093/aje/kwq340

Smolders, J., Damoiseaux, J., Menheere, P., Tervaert, J. W. C., & Hupperts, R. (2009). Association study on two vitamin D receptor gene polymorphisms and vitamin D metabolites in multiple sclerosis. *Annals of the New York Academy of Sciences*, *1173*, 515–520. https://doi.org/10.1111/j.1749-6632.2009.04656.x

Štefanić, M., Karner, I., Glavaš-Obrovac, L., Papić, S., Vrdoljak, D., Levak, G., & Krstonošić, B. (2005). Association of vitamin D receptor gene polymorphism with susceptibility to Graves' disease in Eastern Croatian population: Case-control study. *Croatian Medical Journal*, *46*(4), 639–646.

Štefanić, M., Papić, S., Suver, M., Glavaš-Obrovac, L., & Karner, I. (2008). Association of vitamin D receptor gene 3′-variants with Hashimoto's thyroiditis in the Croatian population. *International Journal of Immunogenetics*, *35*(2), 125–131. https://doi.org/10.1111/j.1744-313X.2008.00748.x

Sucheston, L., Witonsky, D. B., Hastings, D., Yildiz, O., Clark, V. J., di Rienzo, A., & Onel, K. (2011). Natural selection and functional genetic variation in the p53 pathway. *Human Molecular Genetics*, *20*(8), 1502–1508. https://doi.org/10.1093/hmg/ddr028

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, *12*(3), 1–10. https://doi.org/10.1371/journal.pmed.1001779

Sundqvist, E., Bäärnhielm, M., Alfredsson, L., Hillert, J., Olsson, T., & Kockum, I. (2010). Confirmation of association between multiple sclerosis and CYP27B1. *European Journal of Human Genetics : EJHG*, *18*(July), 1349–1352. https://doi.org/10.1038/ejhg.2010.113

Tanaka, A., Nezu, S., Uegaki, S., Kikuchi, K., Shibuya, A., Miyakawa, H., Takahashi, S. ichi, Bianchi, I., Zermiani, P., Podda, M., Ohira, H., Invernizzi, P., & Takikawa, H. (2009). Vitamin D receptor polymorphisms are associated with increased susceptibility to primary biliary cirrhosis in Japanese and Italian populations. *Journal of Hepatology*, *50*(6), 1202–1209. https://doi.org/10.1016/j.jhep.2009.01.015

Tao, C., Simpson, S., Taylor, B. V., & van der Mei, I. (2017). Association between human herpesvirus & human endogenous retrovirus and MS onset & progression. *Journal of the Neurological Sciences*, *372*, 239–249. https://doi.org/10.1016/j.jns.2016.11.060

Tao, C., Simpson, S., Van Der Mei, I., Blizzard, L., Havrdova, E., Horakova, D., Shaygannejad, V., Lugaresi, A., Izquierdo, G., Trojano, M., Duquette, P., Girard, M., Grand'Maison, F., Grammond, P., Alroughani, R., Terzi, M., Oreja-Guevara, C., Sajedi, S. A., Iuliano, G., … Taylor, B. V. (2016). Higher latitude is significantly associated with an earlier age of disease onset in multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry*, *87*(12), 1343–1349. https://doi.org/10.1136/jnnp-2016-314013

Tatomir, A., Talpos-caia, A., & Anselmo, F. (2017). *The complement system as a biomarker of disease activity and response to treatment in multiple sclerosis*.

The GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project The. *Nat Genet*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P., Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., … Cohen, J. A. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *The Lancet Neurology*, *17*(2), 162–173. https://doi.org/10.1016/S1474-4422(17)30470-2

Torabi, M., Green, C., Yu, N., & Marrie, R. A. (2014). Application of three focused cluster detection methods to study geographic variation in the incidence of multiple sclerosis in Manitoba, Canada. *Neuroepidemiology*, *43*(1), 38–48. https://doi.org/10.1159/000365761

Tremlett, H., Zhu, F., Ascherio, A., & Munger, K. L. (2018). Sun exposure over the life course and associations with multiple sclerosis. *Neurology*, 10.1212/WNL.0000000000005257. https://doi.org/10.1212/WNL.0000000000005257

Tschochner, M., Leary, S., Cooper, D., Strautins, K., Chopra, A., Clark, H., Choo, L., Dunn, D., James, I., Carroll, W. M., Kermode, A. G., & Nolan, D. (2016). Identifying patient-specific Epstein-Barr Nuclear Antigen-1 genetic variation and potential autoreactive targets relevant to multiple sclerosis pathogenesis. *PLoS ONE*, *11*(2), 1–22. https://doi.org/10.1371/journal.pone.0147567

Urbach, D., & Moore, J. H. (2011). The spatial dimension in biological data mining. *BioData Mining*, *4*(1), 1–2. https://doi.org/10.1186/1756-0381-4-6

Van Der Mei, I. A. F., Dwyer, t., Blizzard, l., Ponson, A. L., Simmons, R., Taylor, B. V., Butzkueven, H., & Kilpatrick, T. (2003). Past exposure to sun, skin phenotype, and risk of multiple sclerosis: Case-control study. *Bmj*, *327*(7410), 316. https://doi.org/10.1136/bmj.327.7410.316

Vasilopoulos, Y., Sarafidou, T., Kotsa, K., Papadimitriou, M., Goutzelas, Y., Stamatis, C., Bagiatis, V., Tsekmekidou, X., Yovos, J. G., & Mamuris, Z. (2013). VDR TaqI is associated with obesity in the Greek population. *Gene*, *512*(2), 237–239. https://doi.org/10.1016/j.gene.2012.10.044

Vogel, A., Strassburg, C. P., & Manns, M. P. (2002). Genetic association of vitamin D receptor polymorphisms with primary biliary cirrhosis and autoimmune hepatitis. *Hepatology*, *35*(1), 126–131. https://doi.org/10.1053/jhep.2002.30084

Vukusic, S., Van Bockstael, V., Gosselin, S., & Confavreux, C. (2007). Regional variations in the prevalence of multiple sclerosis in French farmers. *Journal of Neurology, Neurosurgery, and Psychiatry*, *78*, 707–709. https://doi.org/10.1136/jnnp.2006.101196

Wallin, M. T., Culpepper, W. J., Coffman, P., Pulaski, S., Maloni, H., Mahan, C. M., Haselkorn, J. K., & Kurtzke, J. F. (2012). The Gulf War era multiple sclerosis cohort: Age and incidence rates by race, sex and service. *Brain*, *135*, 1778–1785. https://doi.org/10.1093/brain/aws099

Wang, Y., & Kasper, L. H. (2014). The role of microbiome in central nervous system disorders. *Brain, Behavior, and Immunity*, *38*, 1–12. https://doi.org/10.1016/j.bbi.2013.12.015

Westerlind, H., Ramanujam, R., Uvehag, D., Kuja-Halkola, R., Boman, M., Bottai, M., Lichtenstein, P., & Hillert, J. (2014). Modest familial risks for multiple sclerosis: A registry-based study of the population of Sweden. *Brain*, *137*(3), 770–778.

https://doi.org/10.1093/brain/awt356

Wheeler, H. E., Aquino-Michaels, K., Gamazon, E. R., Trubetskoy, V. V., Dolan, M. E., Huang, R. S., Cox, N. J., & Im, H. K. (2014). Poly-omic prediction of complex traits: OmicKriging. *Genetic Epidemiology*, *38*(5), 402–415. https://doi.org/10.1002/gepi.21808

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.*

Willer, C. J., Dyment, D. A., Risch, N. J., Sadovnick, A. D., Ebers, G. C., Paty, D. W., Hashimoto, S. A., Devonshire, V., Hooge, J., Oger, J., Metz, L., Warren, S., Hader, W., Nelson, R., Freedman, M., Brunet, D., Paulseth, J., Rice, G., O'Connor, P., … Stefanelli, M. (2003). Twin concordance and sibling recurrence rates in multiple sclerosis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(22), 12877–12882. https://doi.org/10.1073/pnas.1932604100

Wöbke, T. K., Sorg, B. L., & Steinhilber, D. (2014). Vitamin D in inflammatory diseases. *Frontiers in Physiology*, *5 JUL*(July), 1–20. https://doi.org/10.3389/fphys.2014.00244

Wright, K. (2018). *corrgram: Plot a Correlogram.*

Xia, Z., White, C. C., Owen, E. K., Korff, A. Von, Clarkson, S. R., McCabe, C. A., Cimpean, M., Winn, P. A., Hoesing, A., Steele, S. U., Cortese, I. C. M., Chitnis, T., Weiner, H. L., Reich, D. S., Chibnik, L. B., & De Jager, P. L. (2016). Genes and Environment in Multiple Sclerosis project: A platform to investigate multiple sclerosis risk. *Annals of Neurology*, *79*(2), 178–189. https://doi.org/10.1002/ana.24560

Yadav, S. K., Boppana, S., Ito, N., Mindur, J. E., Mathay, M. T., Patel, A., Dhib-Jalbut, S., & Ito, K. (2017). Gut dysbiosis breaks immunological tolerance toward the central nervous system during young adulthood. *Proceedings of the National Academy of*

Sciences, 201615715. https://doi.org/10.1073/pnas.1615715114

Yuen, A. W. C., & Jablonski, N. G. (2010). Vitamin D: In the evolution of human skin colour. *Medical Hypotheses*, *74*(1), 39–44. https://doi.org/10.1016/j.mehy.2009.08.007

Zhao, Z., Zhang, J., Sha, Q., & Hao, H. (2019). Testing gene-environment interactions for rare and/or common variants in sequencing association studies. *BioRxiv*, 796540. https://doi.org/10.1101/796540

9. APPENDIXES

9.1. INPUT DATA FOR APPLICATION 1: MS IN EUROPE

A. MS prevalence and yearly solar irradiation

LONG: longitude.
LAT: latitude.
MS prevalence: MS affected individuals per 100000 people.
Solar irradiation: Long-term average irradiation on horizontal plane (Wh/m^2/day).

LONG	LAT	STUDY REGION	MS PREVALENCE	SOLAR IRRADIATION
22.84	66.33	Överkalix (Sweden)	253	2240
22.85	62.79	Seinäjoki-south (Finland)	219	2420
10.75	59.91	Oslo (Norway)	170	2310
13.51	59.40	County of Värmland (Sweden)	168.3	2470
11.5	64.02	Nord-Trøndelag County (Norway)	165	2250
12.57	55.68	Denmark	154.5	2670
16.52	65.33	Västerbotten (Sweden)	154	2310
11.03	50.98	Erfurt, Thuringia (Germany)	128	2740
21.62	63.1	Vaasa (Finland)	107	2410
25.75	62.24	Central Finland	105	2390
16.37	48.21	Austria	98.8	3140
24.94	60.17	Uusimaa (Finland)	93	2560
11.97	57.71	Gothenburg region (Sweden)	84	2520
9.92	51.54	South Lower Saxony (Germany)	83.5	2670
18.78	69.82	Troms and Finnmark (Norway)	73	1920
26.49	58.06	South Estonia	49.56	2670
-3.19	55.95	South East Scotland	211	2430
-6.25	55.20	Ballymoney, Coleraine, Ballymena, and Moyle districts (North Ireland)	200.5	2500
-8.11	54.65	County Donegal (Ireland)	194.6	2480

LONG	LAT	STUDY REGION	MS PREVALENCE	SOLAR IRRADIATION
-2.97	56.46	Tayside (Scotland)	184	2400
-3.16	56.20	Fife (Scotland)	178	2410
0.97	52.19	Rural Suffolk (England)	153	2710
5.32	60.39	Hordaland County (Norway)	150.8	2200
-4.25	55.86	Glasgow (England)	145	2400
-3.18	51.48	South East Wales	144.6	2800
-2.15	53.60	Rochdale Metropolitan Borough (England)	143	2650
0.12	52.21	North Cambridgeshire (England)	139	2670
-0.13	51.51	Borough of Sutton, London (England)	129	2700
6.18	49.12	Lorraine (France)	123.7	2930
4.36	48.96	Champagne, Ardenne (France)	122.9	3000
-6.46	52.34	County of Wexford (England)	121.2	2730
-2.13	49.21	Jersey, Channel Islands	120.2	3160
-3.53	50.72	Devon (England)	117.6	2940
2.30	49.89	Picardie (France)	116.8	2830
5.04	47.32	Bourgogne (France)	116.2	3200
6.02	47.24	Franche-Comté (France)	115.3	3200
-1.40	50.91	Southampton, South West Hampshire (England)	115	2860
3.06	50.63	Nord-Pas-de-Calais (France)	113.3	2770
7.75	48.57	Alsace (France)	112	2970
-0.14	50.82	Brighton, Mid-downs (England)	111	2800
-1.55	53.80	Leeds (England)	108.7	2550
1.91	47.90	Centre region (France)	103.1	3160
-0.37	49.18	Basse, Normandie (France)	98	3090
-0.99	53.10	Bassetlaw, Nottinghamshire (England)	98	2590
-2.59	49.47	Guernsey (Channel Islands)	95.6	3130
-1.68	48.12	Bretagne (France)	94.7	3200
1.10	49.44	Haute Normandie (France)	94.7	2940

LONG	LAT	STUDY REGION	MS PREVALENCE	SOLAR IRRADIATION
3.72	51.05	Flanders (Belgium)	87.9	2690
-1.55	47.22	Pays de la Loire (France)	86.7	3360
2.35	48.86	Île-de-France (France)	84.3	3050
-21.82	64.13	Iceland	78.5	2130
7.16	62.74	Counties of Møre and Romsdal (Norway)	75.4	2180
15.14	37.84	Town of Linguaglossa, Province of Catania, Sicily (Italy)	197.8	4720
14.06	37.49	City of Caltanissetta, Sicily (Italy)	165.8	4780
12.45	43.94	Republic of San Marino	159.1	3650
9.33	40.32	Area of Barbagia, Nuoro, Sardinia (Italy)	156.4	4560
14.86	45.64	Gorski kotar-Kocevje region (Slovenia and Croatia)	151.9	3350
15.08	45.37	Gorksi kotar (Croatia)	143.5	3440
8.56	40.73	Province of Sassari, Sardinia (Italy)	140.9	4360
14.28	37.6	City of Enna, Sicily (Italy)	120.2	4730
21.73	38.25	Western Greece	119.61	4010
11.62	44.84	Province of Ferrara, Emilia-Romagna region (Italy)	114.28	3500
15.08	37.51	City of Catania, Sicily (Italy)	97.2	4730
13.34	41.64	Province of Frosinone, Lazio region (Italy)	94.4	4000
9.16	45.18	Province of Pavia, Lombardy region (Italy)	94	3530
8.95	44.41	Province of Genoa, Liguria region (Italy)	85	3700
11.88	45.41	Province of Padova, Lombardy region (Italy)	81.4	3460
8.74	41.91	Corse (Italy)	81.1	4170
14.77	40.68	Province of Salerno, Campania region (Italy)	71.62	4130
13.29	38.08	City of Monreale, Sicily (Italy)	71.2	4590
20.14	46.25	Csongrád County (Hungary)	62	3480

LONG	LAT	STUDY REGION	MS PREVALENCE	SOLAR IRRADIATION
13.4	42.35	District of L'Aquila, Abruzzo region (Italy)	55.9	3980
33.38	35.19	Nicosia Metropolitan area (Cyprus)	51.1	4760
32.43	34.77	Paphos district (Cyprus)	44.4	4790
23.32	42.7	Sofia (Bulgaria)	43.05	3510
20.45	44.79	Belgrade (Yugoslavia)	41.5	3650
23.34	42.96	Svoge (Bulgaria)	39.3	3460
24.71	42.89	Trojan (Bulgaria)	39.1	3550
25.87	40.85	Province of Evros (Italy)	38.32	3760
23.55	42.34	Samokov (Bulgaria)	38.1	3710
33.92	35.11	Famagusta district (Cyprus)	34.4	4710
32.86	34.94	Cyprus	33.3	5260
17.81	43.34	Western Herzegovina Canton and Herzegovina-Neretva Canton (Bosnia and Herzegovina)	30.99	4110
16.34	46.31	Varaždin County (Croatia)	29.44	3230
17.59	43.38	Western Herzegovina Canton (Bosnia and Herzegovina)	26.09	4120
24.55	46.53	Mureş County (Romania)	20.97	3710
14.38	35.94	Malta	16.7	4830
7.45	46.95	Berne (Switzerland)	110.43	3220
3.09	45.78	Auvergne (France)	101.5	3510
1.26	45.83	Limousin (France)	93.2	3470
-0.58	44.84	Aquitaine (France)	87.9	3550
1.44	43.60	Midi-Pyrénées (France)	87.4	3720
4.84	45.76	Rhône-Alpes (France)	87.3	3470
0.34	46.58	Poitou-Charentes (France)	85.5	3470
5.37	43.30	Provence-Alpes-Côte d'Azur region (France)	84.1	4200
3.88	43.61	Languedoc-Roussillon region (France)	83.1	4030
-1.11	40.35	Bajo Aragón, Teruel (Spain)	77	4480

LONG	LAT	STUDY REGION	MS PREVALENCE	SOLAR IRRADIATION
-8.54	42.88	Santiago de Compostela, A Coruña (Spain)	72	3700
4.26	39.89	Menorca (Spain)	67.3	4320
-5.66	43.53	Gijón, Asturias (Spain)	65	3280
-1.65	41.35	Sanitary District of Calatayud, Zaragoza (Spain)	58	4250
2.25	41.93	Catalonia, Spain	58	4170
-4.72	41.65	Valladolid (Spain)	53	4250
-4.10	36.78	Sanitary District of Vélez, Málaga (Spain)	53	4570
-4.42	36.72	Málaga (Spain)	52.9	4580
-8.69	39.24	Santarém (Portugal)	46.3	4380
-3.87	40.32	Móstoles, Madrid (Spain)	43.4	4520
7.43	45.74	Valle d'Aosta region (Italy)	39	3770
-0.48	38.7	Alcoy health region, Valencia (Spain)	17.17	4550

B. HLA-DRB1*15:01

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*15:01 frequency
20	41	Albania	0.076
24	52	Belarus Brest Region	0.133
31	52	Belarus Gomel Region	0.121
30	55	Belarus Vitebsk Region	0.185
18.42	43.87	Bosnia and Herzegovina	0.1268
23	42	Bulgaria	0.054
15.55	45.48	Croatia Gorski kotar Region	0.102
15	45	Croatia	0.109
22	40	Greece North	0.068
22.95	40.63	Greece	0.082
19.47	51.75	Poland Lodz	0.146
20.47	44.47	Serbia	0.084
17.02	48.03	Slovakia	0.123
27	50	Ukraine Khmelnytskyi	0.101

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*15:01 frequency
24	50	Ukraine Lvov	0.098
16.35	48.22	Austria	0.132
4.35	50.83	Belgium	0.142
-1.28	52.62	England	0.111
-2.8	54.03	England Lancaster	0.159
-1.53	53.78	England Leeds	0.148
-3	53.4	England Liverpool	0.15
-2.23	53.47	England Manchester	0.163
-1.6	54.97	England Newcastle	0.15
-1.47	53.38	England Sheffield	0.141
18.08	59.55	Finland	0.083
-1.55	47.22	France Grenoble, Nantes, and Rennes	0.119
5	46	France Southeast	0.097
-4.08	47.98	France West Breton	0.183
8.67	50.1	Germany	0.142
-6.75	54.67	Ireland Northern	0.186
9.18	45.47	Italy	0.131
12.5	41.9	Italy Central	0.063
9.15	45.18	Italy North Pavia	0.043
9.15	45.18	Italy North Pavia	0.064
7.7	45.07	Italy North	0.031
9.05	39.23	Italy Sardinia	0.046
10.75	59.93	Norway	0.157
25	70	Norway Sami	0.105
-8.45	40.57	Portugal Aveiro	0.079
-8.05	37.5	Portugal Beja	0.075
-8.33	41.62	Portugal Braga	0.092
-6.83	41.8	Portugal	0.075
-7.47	39.83	Portugal Castelo Branco	0.08
-8	39	Portugal Center	0.14
-8.43	40.18	Portugal Center	0.07

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*15:01 frequency
-8.05	40.22	Portugal Coimbra	0.071
-7.9	38.57	Portugal Evora	0.083
-7.92	37.03	Portugal Faro	0.066
-7.55	40.82	Portugal Guarda	0.073
-8.93	39.75	Portugal Leiria	0.081
-9.18	38.7	Portugal Lisbon	0.083
-8	41	Portugal North	0.12
-7.42	39.32	Portugal Portalegre	0.075
-8.63	41.15	Portugal Porto	0.082
-8.68	39.23	Portugal Santarem	0.08
-8.88	38.53	Portugal Setubal	0.076
-8	37	Portugal South	0.051
-9.18	38.7	Portugal South	0.075
-8.83	41.7	Portugal Viana do Castelo	0.086
-7.8	41.28	Portugal Vila Real	0.073
-7.92	40.67	Portugal Viseu	0.094
2.18	41.37	Spain Barcelona	0.083
2.82	41.98	Spain Catalonia Girona	0.119
-2.17	43.17	Spain Gipuzkoa Basque	0.13
-3.6	37.17	Spain Granada	0.109
1.05	39	Spain Ibiza	0.123
-3.68	40.4	Spain Madrid	0.14
2.98	39.62	Spain Majorca	0.104
4.08	39.97	Spain Minorca	0.123
-5.83	43.35	Spain North	0.146
-5.57	42.6	Spain Northwest	0.096
-7.57	43	Spain Northwest Lugo	0.141
-0.37	39.48	Spain Valencia	0.11
18.07	59.35	Sweden Stockholm	0.156
-0.12	51.5	United Kingdom	0.099
-3.18	51.48	Wales	0.138
-3.18	51.48	Wales	0.143

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*15:01 frequency
41	65	Russia Arkhangelsk	0.198
64	64	Russia Arkhangelsk Pomor	0.179
41	58	Russia Kostroma Region	0.135
47	57	Russia Mari	0.106
37.62	55.75	Russia Moscow	0.133
33.08	68.97	Russia Murmansk Saomi Mixed	0.117
52	67.08	Russia Nenet Mixed	0.045
28	58	Russia Northwest	0.155
30.3	59.97	Russia Northwest	0.125
50.12	53.18	Russia Samara Region	0.141
33	56	Russia Smolensk	0.154
60.33	54.53	Russia South Ural Bashkir	0.092
60.33	54.53	Russia South Ural Russian	0.147
60.33	54.53	Russia South Ural Tatar	0.111
40	59	Russia Vologda	0.144

C. HLA-DRB1*03:01

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*03:01 frequency
20	41	Albania	0.054
24	52	Belarus Brest Region	0.09
31	52	Belarus Gomel Region	0.107
30	55	Belarus Vitebsk Region	0.074
18.42	43.87	Bosnia and Herzegovina	0.0932
23	42	Bulgaria	0.082
15.55	45.48	Croatia Gorski Kotar Region	0.142
15	45	Croatia	0.108
22	40	Greece North	0.074
22.95	40.63	Greece	0.07
19.47	51.75	Poland Lodz	0.107
20.47	44.47	Serbia	0.109
27	50	Ukraine Khmelnytskyi	0.054

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*03:01 frequency
24	50	Ukraine Lvov	0.078
16.35	48.22	Austria	0.118
-1.28	52.62	England	0.153
-2.8	54.03	England Lancaster	0.125
-1.53	53.78	England Leeds	0.146
-3	53.4	England Liverpool	0.158
-2.23	53.47	England Manchester	0.147
-1.6	54.97	England Newcastle	0.153
-1.47	53.38	England Sheffield	0.146
18.08	59.55	Finland	0.028
-1.55	47.22	France Grenoble, Nantes, and Rennes	0.113
5	46	France Southeast	0.097
-4.08	47.98	France West Breton	0.116
8.67	50.1	Germany	0.106
-8.1	55	Ireland Donegal	0.175
-6.75	54.67	Ireland Northern	0.16
-6.45	52	Ireland Wexford	0.185
9.18	45.47	Italy	0.101
12.5	41.9	Italy Central	0.061
9.15	45.18	Italy North Pavia	0.068
9.15	45.18	Italy North Pavia pop 2	0.077
7.7	45.07	Italy North	0.086
9.05	39.23	Italy Sardinia	0.272
10.75	59.93	Norway	0.124
25	70	Norway Sami	0.06
-8.45	40.57	Portugal Aveiro	0.111
-8.05	37.5	Portugal Beja	0.119
-8.33	41.62	Portugal Braga	0.111
-6.83	41.8	Portugal Bragan?	0.118
-7.47	39.83	Portugal Castelo Branco	0.102
-8	39	Portugal Center	0.16

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*03:01 frequency
-8.43	40.18	Portugal Center	0.112
-8.05	40.22	Portugal Coimbra	0.118
-7.9	38.57	Portugal Evora	0.119
-7.92	37.03	Portugal Faro	0.119
-7.55	40.82	Portugal Guarda	0.117
-8.93	39.75	Portugal Leiria	0.124
-9.18	38.7	Portugal Lisbon	0.108
-8	41	Portugal North	0.13
-7.42	39.32	Portugal Portalegre	0.118
-8.63	41.15	Portugal Porto	0.107
-8.68	39.23	Portugal Santarem	0.115
-8.88	38.53	Portugal Setubal	0.11
-8	37	Portugal South	0.143
-9.18	38.7	Portugal South	0.109
-8.83	41.7	Portugal Viana do Castelo	0.084
-7.8	41.28	Portugal Vila Real	0.114
-7.92	40.67	Portugal Viseu	0.104
2.18	41.37	Spain Barcelona	0.121
1.05	39	Spain Ibiza	0.115
-3.68	40.4	Spain Madrid	0.134
2.98	39.62	Spain Majorca	0.132
-4.42	36.72	Spain Malaga	0.137
4.08	39.97	Spain Minorca	0.123
-5.83	43.35	Spain North	0.073
-5.57	42.6	Spain Northwest	0.128
-7.57	43	Spain Northwest Lugo	0.106
-0.37	39.48	Spain Valencia	0.11
18.07	59.35	Sweden Stockholm	0.125
-0.12	51.5	United Kingdom	0.153
-3.18	51.48	Wales	0.166
-3.18	51.48	Wales	0.152
41	65	Russia Arkhangelsk	0.092

LONGITUDE	LATITUDE	COUNTRY/REGION	HLA-DRB1*03:01 frequency
64	64	Russia Arkhangelsk Pomor	0.096
41	58	Russia Kostroma Region	0.123
47	57	Russia Mari	0.052
37.62	55.75	Russia Moscow	0.08
33.08	68.97	Russia Murmansk Saomi Mixed	0.08
52	67.08	Russia Nenet Mixed	0.055
28	58	Russia Northwest	0.087
30.3	59.97	Russia Northwest	0.055
50.12	53.18	Russia Samara Region	0.077
33	56	Russia Smolensk	0.093
60.33	54.53	Russia South Ural Bashkir	0.051
60.33	54.53	Russia South Ural Russian	0.094
60.33	54.53	Russia South Ural Tatar	0.063
40	59	Russia Vologda	0.074

D. Pigmentation SNP rs16891982

LONGITUDE	LATITUDE	COUNTRY	rs16891982 allele G frequency
39	44	Russia	0.759
47	56	Russia	0.843
40	60	Russia	0.944
5.3	60.3	Norway	0.978
18.1	59.3	Sweden	0.957
23.8	61.5	Finland	0.917
24.7	59.4	Estonia	0.988
19.9	50.1	Poland	0.98
12.6	55.7	Denmark	0.97
4.5	51.9	Netherlands	0.97
11.6	48.2	Germany	0.962
9.8	52.6	Germany	0.912
11.4	47.3	Austria	0.903
16.4	48.2	Austria	0.964

LONGITUDE	LATITUDE	COUNTRY	rs16891982 allele G frequency
8.6	47.4	Switzerland	0.941
19.1	47.5	Hungary	0.973
-3	59	United Kingdom	0.984
-5.9	54.6	United Kingdom	0.975
-7.9	53.4	Ireland	0.989
25.4	41.1	Greece	0.891
22.9	40.6	Greece	0.855
10	46	Italy	0.959
11	43	Italy	0.949
10.99	45.44	Italy	0.939
9	40	Italy	0.631
2	49	France	0.852
-1.5	43	Spain	0.939
-3.7	40.4	Spain	0.922
-0.5	38	Spain	0.843
-8.6	42.9	Spain	0.792

E. VDR SNP rs731236

LONG	LAT	REGION/COUNTRY	rs731236 allele T frequency	REFERENCE
25.13	35.33	Heraclion, Crete (Greece)	0.44	(Panierakis et al., 2009)
-6.37	39.47	Cáceres (Spain)	0.65	(García-Martín et al., 2013)
20.85	39.67	Ioannina (Greece)	0.63	(Sioka et al., 2011)
32.86	39.94	Ankara (Turkey)	0.58	(Dogan et al., 2009)
-8.41	40.2	Coimbra (Portugal)	0.60	(Maalej et al., 2008)
-3.7	40.42	Madrid (Spain)	0.57	(Barroso et al., 2008)
-3.7	40.42	Madrid (Spain)	0.61	(Barroso et al., 2008)
22.94	40.65	Thessaloniki (Greece)	0.57	(Vasilopoulos et al., 2013)
22.94	40.65	Thessaloniki (Greece)	0.61	(Emmanouilidou et al., 2015)
28.98	40.996	Istanbul (Turkey)	0.64	(Kurt et al., 2012)

LONG	LAT	REGION/COUNTRY	rs731236 allele T frequency	REFERENCE
12.5	41.9	Rome (Italy)	0.62	(Agliardi et al., 2011)
-1.98	43.32	San Sebastián (Spain)	0.66	(Irizar et al., 2012)
22.27	43.9	Zajecar (Serbia)	0.61	(Ramos-Lopez et al., 2005)
20.45	44.78	Belgrade (Serbia)	0.70	(Kujundzic et al., 2016)
9.19	45.47	Milán (Italy)	0.61	(Tanaka et al., 2009)
9.19	45.47	Milán (Italy)	0.64	(Colombini et al., 2016)
18.7	45.55	Osijek (Croatia)	0.58	(Štefanić et al., 2008)
18.7	45.55	Osijek (Croatia)	0.62	(Rucevic et al., 2012)
15.996	45.81	Zagreb (Croatia)	0.57	(Štefanić et al., 2005)
15.996	45.81	Zagreb (Croatia)	0.58	(Štefanić et al., 2008)
8.54	47.38	Zurich (Germany)	0.59	(Many et al., 2012)
8.67	49.4	Heidelberg (Germany)	0.61	(Abbas et al., 2008)
14.44	50.08	Prague (Czech Republic)	0.63	(Hughes et al., 2011)
8.68	50.11	Frankfurt (Germany)	0.61	(Ramos-Lopez et al., 2005)
5.7	50.85	Maastricht (Netherlands)	0.63	(Smolders et al., 2009)
17.05	51.11	Breslavia (Poland)	0.65	(Łaczmański et al., 2015)
0.12	52.21	Cambridge (United Kingdom)	0.62	(Dunning et al., 1999)
0.12	52.21	Cambridge (United Kingdom)	0.60	(Dunning et al., 1999)
21.01	52.23	Warsaw (Poland)	0.62	(Mostowska et al., 2013)
21.01	52.23	Warsaw (Poland)	0.63	(Horst-Sikorska et al., 2008)
21.01	52.23	Warsaw (Poland)	0.62	(Ramos-Lopez et al., 2005)
9.75	52.39	Hannover (Germany)	0.63	(Vogel et al., 2002)
13.4	52.53	Berlin (Germany)	0.58	(Heine et al., 2013)
-2.97	53.41	Liverpool (United Kingdom)	0.61	(Martin et al., 2010)
14.57	53.42	Szczecin (Poland)	0.66	(Gapska, Scott, Serrano-Fernandez, Huzarski, et al., 2009)

LONG	LAT	REGION/COUNTRY	rs731236 allele T frequency	REFERENCE
14.57	53.42	Szczecin (Poland)	0.64	(Gapska, Scott, Serrano-Fernandez, Mirecka, et al., 2009)
14.57	53.42	Szczecin (Poland)	0.61	(Kempiska-Podhorecka et al., 2012)
9.9	53.55	Hamburg (Germany)	0.61	(Abbas et al., 2008)
-1.55	53.8	Leeds (United Kingdom)	0.59	(Randerson-Moor et al., 2009)
12.59	55.67	Copenhagen (Denmark)	0.58	(Smedby et al., 2011)
55.76	37.62	Moscow (Russia)	0.64	(Karami et al., 2008)
15.62	58.41	Linkoping (Sweden)	0.57	(Lundin et al., 1999)
24.94	60.17	Helsinki (Finland)	0.68	(Sillanpää et al., 2004)

F. CYP27B1 SNP rs12368653

LONGITUDE	LATITUDE	REGION/COUNTRY	rs12368653 allele G frequency
11.6	48.2	Germany	0.5108
9.8	52.6	Germany	0.5473
16.4	48.2	Austria	0.511
24.7	59.4	Estonia	0.543
23.8	61.5	Finland	0.5707
2	46	France	0.517
-1.5	43.5	France	0.479
9.7	45.7	Bergamo (Italy)	0.4352
11.3	44.5	Toscana (Italy)	0.625
9	40	Sardinia (Italy)	0.411
5.3	60.3	Norway	0.5289
39	44	Russia	0.5588
40	61	Russia	0.6
18.1	59.3	Sweden	0.5496
14.1	61	Sweden	0.5718

LONGITUDE	LATITUDE	REGION/COUNTRY	rs12368653 allele G frequency
8.6	47.4	Switzerland	0.511
-3	59	United Kingdom	0.625
-5.9	54.6	United Kingdom	0.53

9.2. SOFTWARE

IMPUTE2 v.2.3.2
B. N. Howie, P. Donnelly, and J. Marchini (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6): e1000529

PLINK v1.90b6.7 64-bit
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81. http://pngu.mgh.harvard.edu/purcell/plink/

R version 3.6.1
R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Complete list of R packages:

- corrplot

Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

- geoR

Paulo J. Ribeiro Jr, Peter J. Diggle, Martin Schlather, Roger Bivand and Brian Ripley (2020). geoR: Analysis of Geostatistical Data. R package version 1.8-1. https://CRAN.R-project.org/package=geoR

- ggbiplot

Vincent Q. Vu (2011). ggbiplot: A ggplot2 based biplot. R package version 0.55. http://github.com/vqv/ggbiplot

- ggplot2

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

- factoextra

Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

- Hmisc

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2019). Hmisc: Harrell Miscellaneous. R package version 4.2-0. https://CRAN.R-project.org/package=Hmisc

- maps

Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2018). maps: Draw Geographical Maps. R package version 3.3.0. https://CRAN.R-project.org/package=maps

- maptools

Roger Bivand and Nicholas Lewin-Koh (2019). maptools: Tools for Handling Spatial Objects. R package version 0.9-9. https://CRAN.R-project.org/package=maptools

- metR

Elio Campitelli (2020). metR: Tools for Easier Analysis of Meteorological Fields. R package version 0.7.0. https://CRAN.R-project.org/package=metR

- Metrics

Ben Hamner and Michael Frasco (2018). Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4. https://CRAN.R-project.org/package=Metrics

- plyr

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

- psych

Revelle, W. (2018) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.8.12

- reshape2

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

- rgl

Daniel Adler, Duncan Murdoch and others (2020). rgl: 3D Visualization Using OpenGL. R package version 0.100.54. https://CRAN.R-project.org/package=rgl

- sm

Bowman, A. W. and Azzalini, A. (2018). R package 'sm': nonparametric smoothing methods (version 2.2-5.6) URL http://www.stats.gla.ac.uk/~adrian/sm

SHAPEIT v2r900

O. Delaneau, J. Marchini, JF. Zagury (2012) A linear complexity phasing method for thousands of genomes. Nat Methods. 9(2):179-81. doi: 10.1038/nmeth.1785

O. Delaneau, JF. Zagury, J. Marchini (2013) Improved whole chromosome phasing for disease and population genetic studies. Nat Methods. 10(1):5-6. doi: 10.1038/nmeth.2307

O. Delaneau, B. Howie, A. Cox, J-F. Zagury, J. Marchini (2013) Haplotype estimation using sequence reads. American Journal of Human Genetics 93 (4) 787-696

J. O'Connell, D. Gurdasani, O. Delaneau, et al. (2014) A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genetics

O. Delaneau, J. Marchini; The 1000 Genomes Project Consortium (2014) Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nature Communications 5 3934