

BACHELOR THESIS  
ARTIFICIAL INTELLIGENCE

**Radboud University**



---

**Analysis of Bayesian quasi-experimental  
designs in geographical contexts to  
evaluate policy effectiveness**

---

*Author:*  
Indy Dolmans  
S1008515

*Supervisor:*  
dr. M. Hinne

*Second reader:*  
dr. P.L. Lanillos



January, 2021

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Bayesian Non-parametric Quasi-experimental Design . . . . .	3
2.2	Gaussian Process Regression . . . . .	5
2.3	Geographical RD . . . . .	6
2.4	Spatial-temporal Regression Discontinuity . . . . .	8
2.5	LATE measures . . . . .	10
<b>3</b>	<b>Simulations</b>	<b>11</b>
3.1	Influence of spread of data points . . . . .	12
3.2	Influence of treatment effect along the border . . . . .	16
3.3	Influence of number of boundary points . . . . .	19
<b>4</b>	<b>Application</b>	<b>20</b>
4.1	Emission of ammonia . . . . .	20
<b>5</b>	<b>Discussion</b>	<b>24</b>
<b>6</b>	<b>References</b>	<b>26</b>
<b>A</b>	<b>Appendix</b>	<b>29</b>

# Analysis of Bayesian quasi-experimental designs in geographical contexts to evaluate policy effectiveness

Indy Dolmans

January 2021

## Abstract

Investigating if a governmental policy implemented in a certain region has the desired effect may be a complex task. Creating randomized controlled trials to evaluate the policy in an experimental setting is not possible. For this matter, Regression Discontinuity (RD) design has been around for a while as an implementation of quasi-experimental designs. This research provides an evaluation of the Bayesian Non-parametric Quasi-experimental Design (BNQD) framework proposed by Hinne, Van Gerven & Ambrogioni (2019) [1], in geographical settings specifically. BNQD is a Bayesian implementation of RD that uses Gaussian Processes to perform quasi-experimental design. The design will be compared to other methods used by researchers in the field of Geographical Regression Discontinuity designs using simulations of data. In order to be able to evaluate the effectiveness of a policy change in one region, the model should be able to capture both spatial data and different time points simultaneously. Combining the two sources of data in a novel framework may rise difficulties, conditions or assumptions that have to be validated in order to achieve a new valid design. The details of the framework are outlined as well as the relation with other quasi-experimental designs. Finally, the framework is applied to a real-life data set of ammonia emissions in the Netherlands to evaluate the far stricter policy applied in Noord-Brabant and Limburg aimed at restricting ammonia emissions from livestock farming. This application shows both the ease with which the framework can be applied and that the emission reduction policy did not have the expected effect.

## 1 Introduction

Researchers in many disciplines all over the world are interested in finding out if their intervention has effect. This intervention can be a medicine, policy or other application which should, directly or indirectly, affect a measurable variable. The approved approach in science to measure the effect of a treatment is by performing randomized controlled trials. It is not always possible to use this form of experimental setup in which samples are randomly distributed over treatment and control group, due to high costs, health risks or physical restrictions. In such cases, Quasi-Experimental Designs (QED) can be utilized to attribute a measured difference exclusively to the intervention, of which several designs exist. One implementation of QED is Regression Discontinuity (RD) [2]. In this approach, the samples are split into a treatment and control group

based on one or multiple assignment variables. Under some assumptions, the partitioning in treatment and control group happens in an as-if random fashion near the threshold of the assignment variable.

The assignment variables can also be taken to represent coordinates on a map (e.g. as longitude and latitude). This causes Regression Discontinuity to evolve into Geographical Regression Discontinuity, or GRD [3]. The assignment variables form a boundary in geographical space, separating two surfaces. Units on one side of the boundary are given treatment, whereas units on the other side are used as a control group. Near the boundary, units appear randomly distributed. This creates a setting in which the difference in outcome variable between the two sides of the border can be directly attributed to the intervention [4].

In a recent paper, Hinne, van Gerven & Ambrogioni (2019) propose a novel framework for performing RD, which they call BNQD, shorthand for Bayesian Non-parametric framework for Quasi-experimental Design [1]. This framework uses Gaussian Process Regression in order to perform Regression Discontinuity [5]. By creating two models, a continuous (no intervention effect is present) and a discontinuous (an effect is present) version, and gathering evidence for either model, posterior evidence in favour of one of the two models states the presence or absence of an effect and is expressed using the Bayes Factor [6].

The BNQD framework can handle two-dimensional data, such as data used in geographical settings, as well as one-dimensional data and can therefore be used in GRD design. This framework, which captures spatial features, may be combined with a dimension of time in order to achieve a novel design which we call spatial-temporal regression discontinuity (STRD). In this design, a function is fitted that takes both two-dimensional geographical data and data from the temporal dimension and returns one outcome variable. This rises possibly new assumptions, conditions and questions on the validity of the design.

This thesis contains an evaluation of the performance of BNQD in geographical settings. The performance will be compared to two other methods used in the field of GRD: the distance-to-border method, which is essentially a reduction of the geographical dimensions to a one-dimensional representation of the distance to the boundary, and the LATE measure method used by Rischard et al (2020) [7], which expresses the treatment effect as a Local Average Treatment Effect to determine the presence of an effect. A comparison metric will be defined with which these methods will be compared. Making use of simulated data, the three methods will be tested and compared. An attempt will be made to add a temporal dimension to GRD to realize a higher-order RD called spatial-temporal RD and implications of the new design will be discussed. Lastly, BNQD will be applied to an ammonia emissions data set in Section 4 to show the practical application of the framework.

## 2 Background

### 2.1 Bayesian Non-parametric Quasi-experimental Design

This thesis builds upon the work of Hinne, Van Gerven & Ambrogioni (2019) to perform quasi-experimental design using the Bayesian framework for non-parametric quasi-

experimental design (BNQD) [1], which will be extended to have a better fit to the peculiarities of two-dimensional settings [3]. The system underlying BNQD will therefore be shortly explained here.

Following the regression discontinuity (RD) design, the data  $D = (x_i, y_i)_i^n$  are split into two disjoint groups: the intervention group, in which samples receive the treatment, and the control group. To which group a sample belongs, is determined by the *assignment variable*. Units with a score of this variable exceeding the predetermined threshold value are assigned to the treatment condition ( $T = 1$ ), whereas the other units are assigned to the control condition ( $T = 0$ ).

In geographical RD, we say that units in a *treated area*  $\mathcal{A}^t$  are compared to units in a *control area*  $\mathcal{A}^c$ , hereby largely adopting the notation laid out by Keele and Titunik (2015) [3]. We can therefore write  $T_i = 1$  if unit  $i$  is in  $\mathcal{A}^t$  and  $T_i = 0$  if unit  $i$  is in  $\mathcal{A}^c$ . The units are divided by a boundary  $\mathcal{B}$ . Adopting the potential outcomes framework, we assert that each unit  $i$  has two potential outcomes,  $Y_{i1}$  and  $Y_{i0}$ , corresponding to  $T_i = 1$  and  $T_i = 0$ , respectively [8]. Effect size  $d$  may then be measured for each unit  $i$  as  $d_i = Y_{i1} - Y_{i0}$ , where the problem arises that it is not possible to observe  $Y_{i1}$  and  $Y_{i0}$  simultaneously. Under some assumptions however, we may be able to identify the treatment effect. The first assumption is the continuity assumption, which is also important for one-dimensional RD but is somewhat more complicated in a two-dimensional geographical context, and holds if, in absence of the considered treatment, units close to the boundary have similar characteristics [3, 9]. The second assumption is the conditional independence assumption, stating that individuals should not be able to self-select into treatment. That is, the attribution of the sample to the treatment or control group should be random. If both assumptions hold, the treatment effect may be estimated using the differences of the model expectations, evaluated at a set of points  $b$  on boundary  $\mathcal{B}$ . These expectations follow from extrapolations of the function to the boundary, computed by BNQD. BNQD then follows the Bayesian paradigm to distinguish an effect by comparison of the continuous model  $M_0$  and discontinuous model  $M_1$ . The continuous model can be regarded as null-model, which indicates that there is no effect present. This model fits one regression function to fit all data points. On the other hand, the discontinuous model is the alternative model indicating that the treatment has effect. The model fits the data in the intervention and control group separately. The model then quantifies the effect size as a Gaussian distribution, evaluated at a discrete set of boundary points  $b \in \mathcal{B}$ :

$$\begin{aligned}
 d_{b|Y} &\sim \mathcal{N}(\mu_{b|Y}, \Sigma_{b|Y}), \text{ with} \\
 \mu_{b|Y} &= \mu_{b|T} - \mu_{b|C} \\
 \Sigma_{b|Y} &= \Sigma_{b|T} + \Sigma_{b|C},
 \end{aligned} \tag{1}$$

where  $T$  and  $C$  indicate the predictions obtained using data in the treatment, respectively control region and  $Y = T \cup C$  [1, 7].

A key property of BNQD is the use of Bayes Factors to quantify evidence in favour of either model, essentially capturing evidence for the presence or absence of an effect [6]. This is different than standard hypothesis testing, which only gathers evidence to *reject* the null-hypothesis that states no effect is present. Bayes Factors directly quantify evidence, with a larger BF indicating that there is a stronger confidence in the

truth of the outcome. Bayes Factors are expressed as the ratio likelihood of data under the alternative (discontinuous) model to the data likelihood under the null (continuous) model:

$$BF_{10} = \frac{p(D | M_1)}{p(D | M_0)}. \quad (2)$$

Once we know that an effect is present, we may be interested in discovering the size of the effect. The two models can also give an estimate of the discontinuity  $d$ . The use of Bayesian Model Averaging (BMA) is a distinguishing feature of BNQD [10]. BMA is used to give a weighted average of the estimation of the true effect size, weighing the effect size estimates by their respective model likelihoods. The effect size estimate is then given by:

$$p(d | D) = \sum_{j=0,1} p(d | D, M_j) p(M_j | D), \quad (3)$$

yielding the sum of the estimated effect size  $d$  for both the continuous and discontinuous model, weighted by their model posterior. The continuous model  $M_0$  predicts an *absence* of discontinuity along the boundary, i.e. the distribution is simply a spike at  $d = 0$ , whereas the effect size estimate according to the discontinuous model  $M_1$  follows a Gaussian distribution.

## 2.2 Gaussian Process Regression

Studies involved in Regression Discontinuity design generally have focused on fitting two polynomials, one to the samples in the treatment group and one to the control group samples, to the data and evaluating the difference at the border. In this thesis, we exploit Gaussian Processes (GP) to fit to the data [5]. This technique has also been used in the paper by Branson et al. (2019) and in their follow-up paper by Rischard et al. (2020) [7, 11]. The advantage of GPs lies in that there are no prior assumptions made about the form of the function that generates the data. We therefore call Gaussian Processes *non-parametric models*. By not placing explicit assumptions on the form of the data-generating function, the GP can be interpreted as a combination of an infinite number of Gaussian distributions. We only assume that the function follows a GP prior:

$$f(x; \theta) \sim \mathcal{GP}(\mu(x), k(x, x')) \quad (4)$$

A GP is described by a *mean function*  $\mu(x)$  and a *covariance function*  $k(x, x'; \theta)$ , also referred to as the *kernel*. Generally, the mean is fixed at 0 for simplicity. This means that the GP is exclusively represented by the covariance function  $k(x, x')$  and its hyperparameters  $\theta$ , such as the length scale parameter  $\ell$  and the GP output variance  $\sigma^2$ . The covariance function computes the similarity between two points of the function,  $f(x)$  and  $f(x')$ . This metric can be computed in several ways and is described by the choice of a certain kernel. Prior knowledge about the shape of  $f$  may be incorporated to make a kernel choice. Many different kernels exist, each exhibiting certain characteristic behaviour. For example, a polynomial kernel exists, able to capture a polynomial trend, or a Matérn kernel, characterized by its smooth adaptation to the true data-generating function. The kernel has a large impact on the behaviour of the functions that are fit to the data. Unreliable results may be produced if a kernel is chosen with a bad fit to the data and may ultimately lead to wrong conclusions. By choosing a kernel, a prior is set

on the structure of the data. Choosing the right covariance function is hence crucial, but this choice may not be straightforward. In a real-life application, it will hardly ever be the case that the chosen covariance function is the true function underlying the data generation. Instead, we assume that the function is sufficiently flexible to provide a smooth fit to the data.

Equivalent to averaging the posterior estimates for the continuous and discontinuous model by weighing them according to their posterior likelihoods, Bayesian Model Averaging provides a way to integrate the estimates of the different covariance functions to compute a log Bayes Factor [10]:

$$BF_{10}^{total} = \frac{p(D | M_1)}{p(D | M_0)} = \frac{\sum_{k \in K} p(D | k) p(k | M_1)}{\sum_{k \in K} p(D | k) p(k | M_0)}. \quad (5)$$

In this paper, we make use of three different covariance functions as similarity measures in the Gaussian Process. The simplest function considered is the linear kernel  $K_{LIN}$ . This is a non-stationary kernel, sensitive to the difference in mean and slope between the intervention and control group [12]. The next kernel is the stationary Matérn kernel  $K_{MAT}$ , which has a parameter  $\nu$  that controls the smoothness of the function. In this thesis, we specify  $\nu = 1.5$  to make the kernel able to detect discontinuities in once differentiable functions, which yields a flexible and smooth fit [5]. The last function considered is the exponential kernel  $K_{EXP}$ , which is a special case of the Matérn function where  $\nu = 1/2$ . This is the simplest stationary kernel and is often used in spatial statistics and is for that matter also included in this research [13]. The exact definitions of these kernels can be found in Appendix A.

### 2.3 Geographical RD

Gaussian Processes may also be useful in the context of geographical RD (GRD). The BNQD framework works with GPs and is built originally to find the effect of a treatment in one-dimensional data, with some parts extended for use in multidimensional cases. In this work, we extend BNQD to be able to use and evaluate BNQD in geographical settings.

One attempt to carrying out RD in geographical settings has been used widely in the field of GRD and entails reducing the two-dimensional problem to a 1D-RD design [14–17]. Instead of keeping the two assignment variables (e.g., longitude and latitude) where the split in treatment and control group is determined by a border  $\mathcal{B}$ , the assignment variable is taken to be the closest distance to the boundary  $\mathcal{B}$ . Here, a negative distance indicates that units are in the control group. In Figure 1, this reduction is shown by mapping the data points from the 3D figure onto a 2D plot. By carrying out this reduction, it becomes possible to perform standard RD on the remaining one-dimensional data to measure an effect. However, this method fails to capture spatial variation. Using the naive distance to the border as assignment variable, there is no way to recover from the problem of spatial heterogeneity, i.e. varying treatment effects *along* the boundary. According to Keele and Titiunik, spatial variation of the treatment effect does not constitute a problem for the validity of the RD design, as long as the assumptions are met, since the design evaluates the effect at a set of border points independently (see Keele & Titiunik (2015), Appendix D).

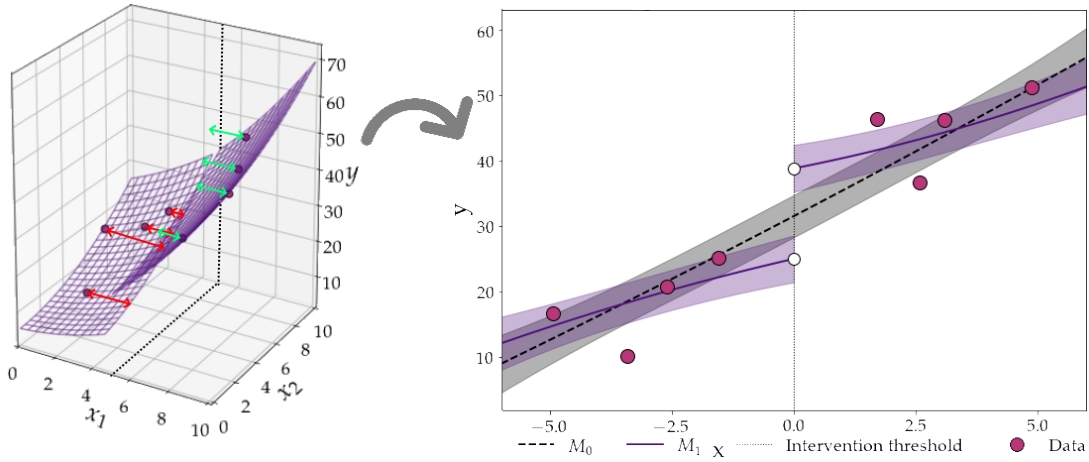


Figure 1: Example of a reduction of GRD with two assignment variables and an outcome variable (left) to 1D-RDD with one assignment variable and one outcome variable (right). Data points are mapped from 3D-GRD plot onto the 2D plot by using the absolute distance to the border. The length of the colored arrows indicate the distance to the border per data point, which is used as the x-axis in the 1D-RD plot. A red arrow indicates that the unit is in the control group, whereas a green arrow indicates a treated unit.

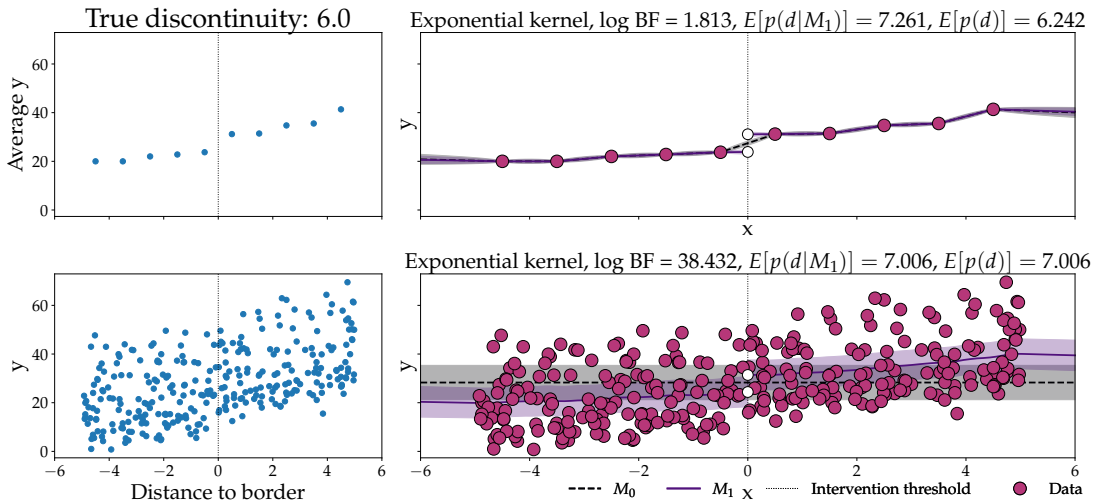


Figure 2: Geographical RD reduced to 1D-RDD using distance to border, where the border lies at 0. In the top plot, the data points with a similar distance to the border are averaged, whereas the plot in the bottom fits the GP to all available data points.

Two versions of this method exist and are both used in the literature. The first version applies the distance-to-border method directly to all data, meaning that the function fit is applied on all separate data points. The second version entails averaging the values of the data points with a similar distance to the border and fitting a function on these bin averages. Figure 2 shows these two options and fits a GP with exponential kernel to the data.

The technique laid out by Rischard et al. fits the data using Gaussian Process Regression, after which the two functions are evaluated at the boundary to find an

estimate of the treatment effect. The posterior distribution is then evaluated at several points on the boundary, which Rischard calls to be ‘sentinel points’. Up to this point, the method is identical to BNQD.

However, there are some important differences between the two methods. A first difference is that the sentinel point method evaluates the treatment effect locally at each sentinel point and calculates a Local Average Treatment Effect (LATE). Building on the effect size estimates per sentinel point, LATE uses a weighting scheme to calculate an average effect size. Using Equation 1, the LATE may be approximated accordingly by incorporating a weighting scheme:

$$d \approx \frac{\sum_{b \in \mathcal{B}} w(b) d_{b|Y}}{\sum_{b \in \mathcal{B}} w(b)}, \quad (6)$$

where  $w(b)$  is a vector of weights evaluated at the sentinel points  $b$ . Several ways to weigh the effect sizes at the different sentinel points are proposed by Rischard, which will be described in detail in Section 2.5. In contrast, BNQD stops after estimation of the effect size per sentinel and uses the evaluation of the effect at these sentinel points to obtain a treatment effect curve. The advantage of not reducing the estimates to one number is that potential spatial variation will be captured, in case the treatment effect along the boundary is heterogeneous. Another difference between the sentinel point method and BNQD is the utilization of Bayesian Model Averaging to define the importance of the continuous and discontinuous model and determining the weight of each kernel [10]. This reduces the overconfidence in the effect presented in  $M_1$  by explicitly taking the continuous model, that is, the probability that no effect is present ( $d = 0$ ), into account. This only occurs when BNQD is not strongly favouring either model in which case weighing according to the model posterior lowers the confidence in the alternative model.

In this thesis, both the sentinel point method and the distance-to-border method have been implemented in Python and will be compared to BNQD. The performance of each method will be evaluated using the Root-Mean-Squared Error (RMSE) [1, 11]. This commonly used error metric computes the distance between the predicted effect size and the true effect size for each of the sentinel points. The LATE measure and the distance-to-border method express the estimation of the treatment effect in one value, where BNQD yields a Gaussian distribution of the effect size for each of the sentinel points. The error produced by the distance to border and LATE measure method is hence simply the difference between the true discontinuity and the predicted discontinuity, whereas the BNQD error will be computed by taking the root of the mean squared difference between the true discontinuity and the predicted discontinuity, evaluated at every  $b \in \mathcal{B}$ .

## 2.4 Spatial-temporal Regression Discontinuity

For certain applications, Geographical Regression Discontinuity (GRD) may be too limited in its flexibility. This may occur in complex situations, for example when viewing a trend over time split by one intervention. By incorporating data from the dimension of time, the GRD design evolves into spatial-temporal RD (STRD), using data points measured at different geographical locations and different points in time. Central in the

## Continuous, exponential kernel    Discontinuous, exponential kernel

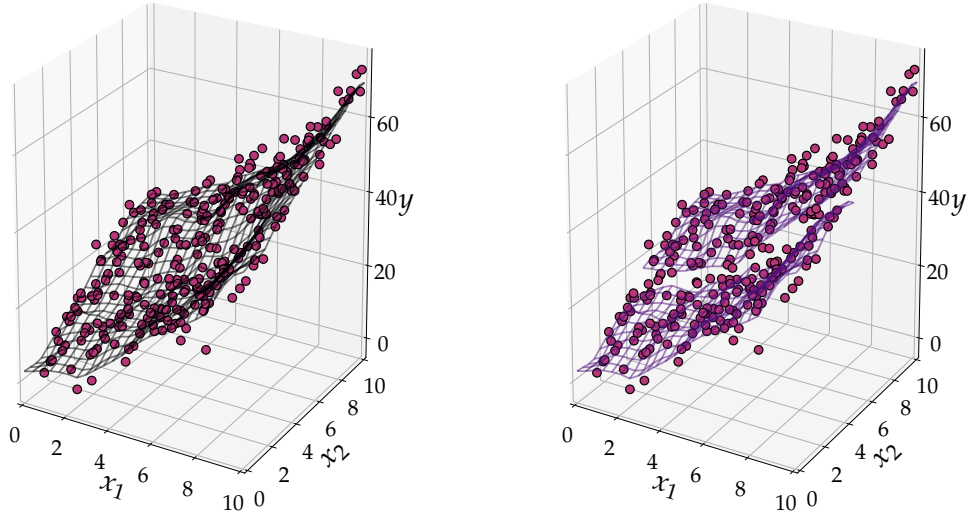


Figure 3: Simulation of an outcome variable on two-dimensional data, shown in a 3D plot. The plot on the left shows one fitted surface for all data points, whereas in the right plot two functions are fitted, split by the assignment boundary. The fitted functions are Gaussian Processes with an exponential kernel.

design is the outcome variable, dependent on both the spatial location and the point in time the measurement has been made. Following this setup, a policy implemented in a certain geographical area may be evaluated if data before and after treatment is available for both the area in which the policy is implemented (treatment area) and an adjacent control area. The boundary consists of the physical border separating the treatment and control area and of the moment in time the policy is implemented, making the design similar to a Difference-in-Differences design [18]. Spatial-temporal RD is similar to a spatial Difference-in-Difference design (S-DiD) in that the difference is used between the differences over time between treated and control units to estimate an average treatment effect. The standard DiD design uses two time periods, one before and one after the intervention, but accommodating additional time periods is straightforward [19]. S-DiD builds on standard DiD but validation of the design requires one important additional assumption, which is the stable unit treatment value assumption, or SUTVA [20]. Causal effects cannot be established in case of violation of this assumption, referred to as ‘social interaction’. This occurs when outcomes of one unit depend on the treatment status of other units, for which reason it is insufficient to establish a causal effect based only on the unit’s own treatment status. Spatial interaction in the treatment is not invalidating the design, but spatial interaction in the responses is a violation of SUTVA [9]. This assumption is an extension of the independence assumption used in GRD design.

Two versions of the STRD design can be developed. The easiest version is performed by eliminating the dimension of time by reducing the outcome variable at different time points to one value. This only works if the time dimension consists of exactly two time periods: one before and one after the intervention occurred. The difference between

the values of the outcome variable before and after the intervention point is then used to perform RD with the coordinates of each data point as the assignment variable. The percentage increase or decrease may also be used instead of the absolute difference as outcome variable.

The reduction to a two-dimensional GRD will not suffice when provided with multiple time points pre- and post-intervention. In this case, the temporal dimension may be represented using an additional assignment variable. This three-dimensional RD design then is a function of three assignment variables, two on the samples' spatial location and one on its point in time, mapping to a single outcome variable. STRD then takes place in the same manner as GRD, as it is straightforward to extend RD so that it can take assignment variables in dimensions of higher order [21].

STRD has similarities to interrupted time series (ITS), which is a quasi-experimental design based on only a time dimension [22]. In contrast to Regression Discontinuity designs, that mainly focus at the cut-off point (the 'boundary'), ITS observes a change in trend by performing a regression and observing a change in the slope from pre-intervention to post-intervention, as well as the level change at the cutoff [23]. ITS also has a set of assumptions that need to be validated, including SUTVA. These requirements are set out in Bernal, Cummins, & Gasparrini (2017) [24]. According to this paper, a strict differentiation of the boundary between pre- and post-intervention points is necessary. Also, the data used in ITS setups need to cover both pre- and post-intervention time periods. Having more data available is beneficial to the power of the conclusions that can be drawn, as confounding variables play a smaller role or may even be excluded. As the temporal dimension is similar in ITS and STRD, these assumptions are important for STRD as well. The framework will be used in an application to evaluate effectiveness of a policy on ammonia emissions in Section 4.

## 2.5 LATE measures

In their paper, Rischard et al. (2020) describe six possible weighting schemes to obtain an average treatment effect (LATE). For this thesis, four of those have been implemented and will be tested:

- Uniform weighted  $w_{UNIF}$ : All weights are set to 1.
- Density weighted  $w_{DENSITY}$ : The weight of every sentinel point is equal to the estimation of density of data points near that point.
- Inverse variance weighted  $w_{INVVAR}$ : The weight of every sentinel point is proportional to the inverse of the variance at that point. This gives high weights to sentinel points where variance of the treatment effect is small and lower weights if the variance is larger.
- Projected land measure  $w_{GEOM}$ : The weight of every sentinel point is equal to the sum of the measured densities for each of the grid points which have their projection to that border point. Grid points are created around the border and data density is estimated at each point, after which the point is projected onto the border. This method is similar to the density weighted LATE, with as main difference the estimation of data density:  $w_{GEOM}$  takes the sum of several estimates where  $w_{DENSITY}$  relies on one estimation taken at the border.

The first three methods do not have any additional parameters. The last measure takes two parameters: the size of the grid, specified as the maximum distance from the border the grid should be created, and the total number of grid points to be created. The more grid points are created, i.e. the spacing between the grid points decreases, the more precise the estimate of the density and hence a lower variance. This comes at the cost of the time it takes to project each of the grid points onto the border. There is thus a trade-off between variance and efficiency. For a more extensive explanation and discussion of the advantages and disadvantages of each weighting procedure, we refer to Rischard et al (2020), Section 3 [7].

By applying the selected weights in Equation 6, we get an estimation of the discontinuity  $d$ . Since the predicted discontinuity  $d_{b|Y}$  is the difference between two Gaussians  $d_{b|T}$  and  $d_{b|C}$ , the effect itself is normally distributed as well. By summing each Gaussian multiplied with a weighting scalar, the LATE remains a Gaussian distribution. The estimated average treatment effect is hence described by its mean  $\mu_d$  and covariance  $\Sigma_d$ . In the next section, the LATE measures are tested using several simulation cases.

### 3 Simulations

In this section, the performance of BNQD will be evaluated using simulations of data points. Each simulation contains two independent variables as representation of a geographical space. A third variable is simulated using a second-order polynomial that is affected by both variables. The data-generating procedure is described using the following process:

$$\begin{aligned}
 x_{ij} &\sim \mathcal{U}(0, 10), \quad i = 1 \dots N, \quad j \in \{1, 2\} \\
 b &= 5.0 \\
 \sigma &= 1.0 \\
 f(x_i) &= 3.0 + 1.1 * x_{i1} + 1.8 * x_{i2} + 0.3 * x_{i1}^2 + d[x_{i2} \geq b] \\
 y_i &\sim \mathcal{N}(f(x_i), \sigma^2),
 \end{aligned} \tag{7}$$

where  $N$  is the number of data points to be sampled and  $x$  is a two-dimensional variable with its values (in the baseline simulation) drawn from a uniform distribution. This distribution will be altered in later simulations. In this setup, the border is always a straight line dependent on only  $x_2$ .

BNQD will be compared to the LATE method described by Rischard et al. (2020) and to the distance-to-border method described in several papers, with its first occurrence in the paper by Black (1999) [7, 14]. The three GRD methods will be compared in several simulations of data, in which two properties of the data will be modified. The first property that will be varied is the density of data points. The density may be constant, varying along the border or varying between the treatment and control group. The second property that will be modified is the true treatment effect along the border. The treatment effect does not have to be constant, but may be increasing, decreasing or shifting discontinuously along the border. The three methods will be applied to a set of covariance functions to view the impact of kernel choice.

Using the root of the mean-squared error (RMSE), the predictions of each of the methods will be compared to the true treatment effect at every boundary point  $b \in \mathcal{B}$ . The square root is taken from the mean error over the boundary points, giving a measure of performance for the three methods, with which they may be compared to one another. As the prediction of the effect is constant in the LATE measures and the distance to border method, the error is simply the difference of the prediction and the true effect evaluated at one point. For BNQD, the RMSE is given by the following formula:

$$RMSE = \sqrt{\frac{1}{N} \sum_b (\text{pred}_b - d_b)^2}, \quad b \in \mathcal{B}, \quad (8)$$

where  $N$  is the number of boundary points  $b$ ,  $\text{pred}_b$  is the BNQD prediction of the treatment effect at boundary point  $b$  and  $d_b$  is the true treatment effect at that point.

The effect of several properties of the data on the estimation of the treatment effect will be investigated and for each of the models the strengths and weaknesses will be demonstrated. For example, since BNQD is the only method that returns a distribution for each boundary point, this method is expected to outperform the single-number LATE measure and distance to border method in simulations with a fluctuating treatment effect. In cases of constant treatment effect however, the discontinuity may be predicted best by the averaged values produced by the LATE and distance to border approaches. As the data are drawn from random distributions, the results may vary per run. To account for fluctuations over different runs, the resulting RMSEs are averaged over 100 runs. This gives a more precise estimate of the average error. Next to the RMSE averages we show the Standard Error of the Mean (SEM), which is a measure of error from the measurements [25]. It is defined as the standard deviation divided by the square root of the number of runs and tells how precise the estimate is.

### 3.1 Influence of spread of data points

In a real-life application, it is unlikely that units are spread evenly across sample space. Samples may be abundant on only one side of the boundary, i.e. in either treatment or control group. It may also be that parts along the boundary have a very high data density, for example when the boundary crosses a city, or a density that approaches zero, when the boundary crosses a lake, forest or highway. The impact of the data density on the different measures will be investigated in this section, where we will vary data density across sample space in four simulations. The first simulation adopts a constant density, an even spread of data points, which is used as a baseline. The following simulation explores the effect of gradually increasing the data density along the border. The third simulation considers a high density in the treatment area and a low density in the control area. The last simulation studies the effects of a high data density both in the left part of the treatment area and in the right part of the control group. This last simulation is taken from the paper written by Rischard et al (2020) where they used this setup when examining the outcomes for the several LATE measures [7].

In the top four plots in Figure 4, the data points generated by the model in Equation 7 are shown for one of the runs of each of the simulation specifications, with the background shaded according to the estimated data density. Below each plot, the

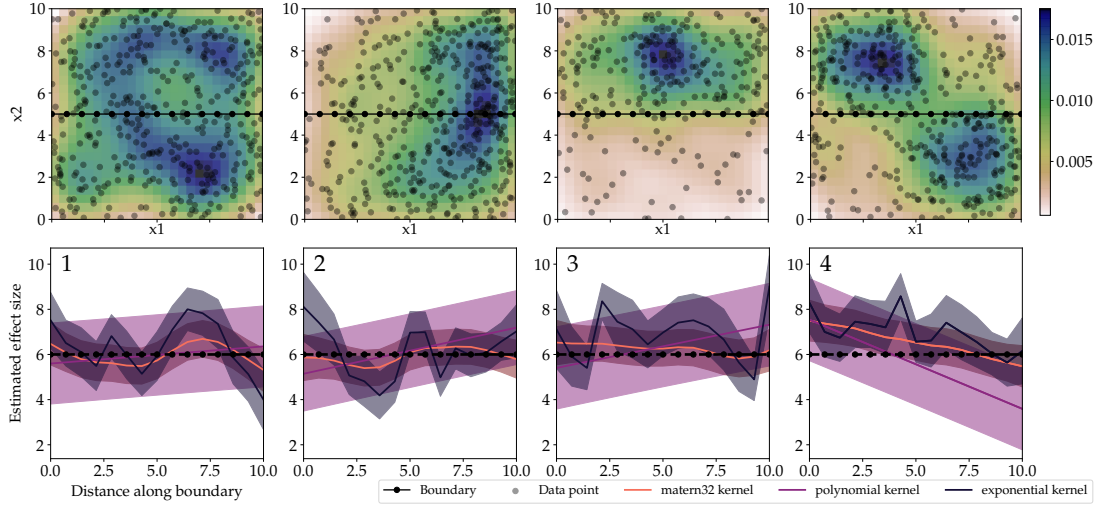


Figure 4: Data points generated for one run of each of the four simulation specifications. Data are overlaid with the estimated data density. Below each plot, the treatment effect predictions for each of the kernels is shown, together with uncertainty.

prediction along the boundary made by the kernels is shown. For each simulation we set a constant treatment effect of  $d = 6.0$ . The predictions made by the three methods will be evaluated on a set of 15 equally spaced points, sampled along the boundary. The results from each of the four density specifications are saved per simulation and per kernel. These are shown in Table 1, which lists the average RMSE and standard error.

As can be seen from this table, the Matérn kernel appears to perform best in all four simulations. In every simulation, using either kernel, the LATE measures outperform BNQD and the reduced distance-to-border method. The density-weighted LATE performs best in all but the third simulation, where the inverse-variance weighted LATE performs slightly better. All LATE measures are generally close in RMSE. The Matérn kernel performs best consistently, whereas the RMSE of the Linear and Exponential kernel is close to each other. The local averaged treatment effects are seemingly closer to the true effect than is the list of predictions produced by BNQD, reflected by a lower RMSE. The BNQD BMA average produces results similar to BNQD, which is expected as the relatively high treatment effect ( $d = 6.0$ ) causes BNQD to have strong beliefs in the discontinuous model. The BMA hence barely takes the continuous model (stating that  $d = 0.0$ ) into account. The BMA average produces an error equal to or larger than BNQD consistently, which is not surprising as the BMA average explicitly lowers the estimated effect size, even though we are here searching for the true discontinuity. The estimation of the reduced 1D-RDD yields large errors, up to six times the error produced by BNQD and even up to 12 (in Simulation 3) times higher than some of the LATE measures. When looking at the standard errors (SEM) per method, listed in the table within parentheses, some interesting features show up. Seen per kernel, the Matern kernel consistently produces the lowest SEM. BNQD and the LATE measures produce similar values, while the distance-to-border method yields variability of the measurements of up to ten times larger than the BNQD and LATE measures. As the

	Simulation 1			Simulation 2		
	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$
BNQD	0.621 (0.031)	1.078 (0.037)	0.553 (0.018)	0.637 (0.035)	1.154 (0.040)	0.599 (0.022)
BNQD BMA	0.621 (0.031)	1.078 (0.037)	0.553 (0.018)	0.637 (0.035)	1.153 (0.040)	0.599 (0.022)
$\mathcal{W}_{UNIF}$	0.462 (0.034)	0.520 (0.043)	0.282 (0.019)	0.463 (0.032)	0.629 (0.032)	0.301 (0.021)
$\mathcal{W}_{DENSITY}$	<b>0.454</b> (0.033)	0.492 (0.039)	<b>0.275</b> (0.020)	<b>0.393</b> (0.027)	<b>0.475</b> (0.027)	<b>0.250</b> (0.019)
$\mathcal{W}_{INVVAR}$	0.462 (0.034)	<b>0.484</b> (0.039)	0.278 (0.019)	0.461 (0.032)	0.524 (0.032)	0.283 (0.020)
$\mathcal{W}_{GEOM}$	0.458 (0.033)	0.499 (0.041)	0.277 (0.020)	0.389 (0.027)	0.498 (0.033)	0.249 (0.020)
1D-RDD	3.038 (0.209)	3.558 (0.237)	2.617 (0.203)	2.883 (0.196)	2.900 (0.212)	2.372 (0.184)
1D-RDD avg	2.973 (0.213)	2.962 (0.216)	2.749 (0.220)	2.876 (0.190)	2.631 (0.191)	2.431 (0.194)
	Simulation 3			Simulation 4		
	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$
BNQD	1.325 (0.040)	1.047 (0.035)	0.614 (0.027)	0.963 (0.046)	1.773 (0.086)	0.833 (0.030)
BNQD BMA	1.325 (0.040)	1.219 (0.087)	0.614 (0.027)	0.982 (0.047)	3.025 (0.018)	0.833 (0.030)
$\mathcal{W}_{UNIF}$	<b>0.479</b> (0.036)	0.435 (0.037)	0.298 (0.027)	0.676 (0.046)	0.991 (0.074)	0.438 (0.033)
$\mathcal{W}_{DENSITY}$	0.489 (0.036)	0.404 (0.036)	0.295 (0.025)	0.676 (0.045)	0.875 (0.064)	<b>0.419</b> (0.030)
$\mathcal{W}_{INVVAR}$	<b>0.479</b> (0.036)	<b>0.400</b> (0.035)	<b>0.291</b> (0.026)	0.676 (0.046)	<b>0.840</b> (0.059)	0.423 (0.032)
$\mathcal{W}_{GEOM}$	0.488 (0.036)	0.419 (0.037)	0.295 (0.027)	<b>0.671</b> (0.046)	0.898 (0.064)	0.422 (0.031)
1D-RDD	6.770 (0.187)	6.099 (0.107)	6.231 (0.117)	4.423 (0.287)	5.680 (0.276)	4.890 (0.283)
1D-RDD avg	6.913 (0.217)	6.514 (0.265)	6.842 (0.300)	4.220 (0.296)	5.218 (0.329)	5.090 (0.345)

Table 1: Results of data-density varying simulations. For every simulation, the average RMSE is shown for every tested method on each of the three kernels. The SEM is displayed within parentheses, the lowest error per simulation per kernel is shown in bold.

fit to the data in this method is rather poor, the resulting error has a wide spread of values, causing the standard error to be large.

The errors produced per LATE measure across the four simulations are approximately similar, from which we may conclude that the density of data points is not of considerable importance to the effect size predictions by the LATE averages in these simulations. BNQD is a bit susceptible to the density of data points, with the RMSE in the third simulation setup being twice as large as the error in the first two simulations.

The distance to border method produces varying errors per simulations, indicating that the method is highly susceptible to the density of data points.

The four simulations are not representative for real-world situations, although they exhibit certain interesting characteristics. The properties of these simulations, such as an imbalanced density between treatment and control group, can be combined to obtain a more realistic simulation. For this, we make use of a Gaussian Process to simulate a ‘realistic landscape’, where cities (areas with a high data density), villages (areas with an intermediate density) and forests (areas with a low data density) may exist. The landscape is modeled using a GP with a Rational Quadratic (RQ) kernel (see Appendix A for the mathematical formula) with a low length scale parameter  $\ell$ . A low length scale (e.g.,  $\ell = 0.1$ ) makes the function less smooth and more wiggly, which is exactly what is desired.

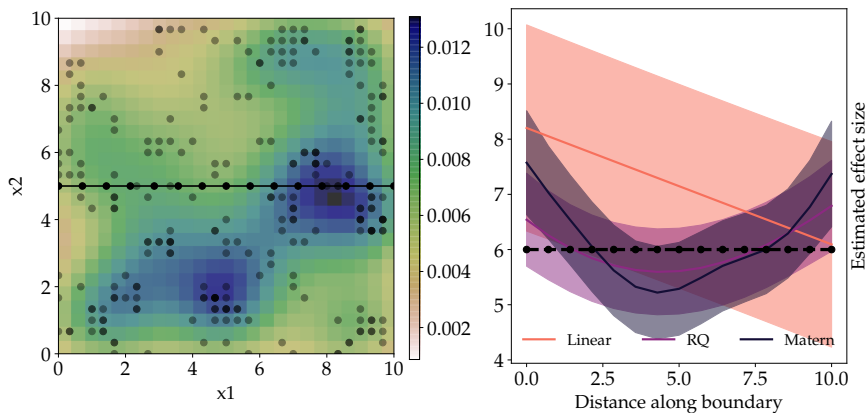


Figure 5: Simulation of a ‘realistic landscape’. Left: generated data points, overlaid with estimated data density. Right: effect size predictions along the boundary for each of the kernels, together with the uncertainty.

The simulated data points from a GP with an RQ kernel are shown for one run in the left plot of Figure 5. The right plot shows the predictions made by a Linear, RQ and Matérn ( $\nu = 3/2$ ) kernel. The data points show a practical interpretation: areas with a high density can be thought of as cities, whereas lower-density areas can be explained as being rural areas. The right plot shows that the estimations of the effect size are not constant along the boundary. The three kernels predict an effect size above the true effect size in the left part of sample space, although the estimates are decreasing for all three kernels and are approaching the true effect size (black dotted line at  $d = 6.0$ ). The linear kernel starts with a prediction fairly above the true effect size, but has a decreasing trend along the boundary. At the other side of sample space, the kernel’s prediction of effect size converges with the true effect size. Both RQ and Matérn have a decreasing estimate along the first part of the boundary, even dropping the estimate slightly below the true effect size, after which an increase is visible again. As the data points are generated using an RQ kernel ( $\ell = 0.1$ ), it is interesting to see if this kernel can recover the function using only a fairly small amount of data points ( $n = 250$ ). The effect size estimation along the boundary is a rather smooth line around the true effect size, indicating a good fit and recovery of the effect size. The Matérn follows the same path, although a bit less smooth. As in the first and last part of sample

space the two kernels overestimate the effect size, while in the middle part there is an underestimation of the effect, the average treatment effect along the boundary would be a good approximation. This will then also be reflected by the LATE measures, that indicate exactly that, as seen in Table 2.

	BNQD	BNQD BMA	$\mathcal{W}_{UNIF}$	$\mathcal{W}_{DENSITY}$	$\mathcal{W}_{INVVAR}$	$\mathcal{W}_{GEOM}$	1D-RDD	1D-RDD avg
Linear	1.544 (0.074)	1.547 (0.075)	1.088 (0.074)	<b>1.070</b> (0.077)	1.086 (0.074)	1.084 (0.072)	6.220 (0.442)	6.588 (0.468)
RQ	0.638 (0.036)	0.638 (0.036)	0.369 (0.033)	<b>0.324</b> (0.029)	0.353 (0.032)	0.331 (0.030)	5.648 (0.381)	6.847 (0.476)
Matern	0.792 (0.036)	0.805 (0.040)	0.382 (0.031)	<b>0.329</b> (0.027)	0.344 (0.028)	0.333 (0.026)	5.904 (0.354)	6.583 (0.444)

Table 2: Results from the ‘realistic landscape’ simulation. The average RMSE over 100 runs is shown per kernel, per method. Within parentheses, SEM is displayed and the lowest RMSE per kernel is shown in bold.

From the results in Table 2 it appears that the Rational Quadratic and Matern kernel have a similar performance, based on the RMSE between the predictions and the true effect size. Only for the BNQD-based methods, using the effect curve rather than an average, the RQ kernel performs slightly better, which is also visible in the right plot of Figure 5. On average, the kernels show a similar performance visually, yet the Matérn kernel shows more wiggles and is less smoothly centered around the true effect size. Compared to the four other simulations with a varying data density, the performance of every method is worse than in the baseline simulation (Simulation 1), but rather similar to the errors produced in the other three simulations. From these results, we may conclude that the behaviour of the methods is affected when using a more realistic data density, compared to keeping the data density artificially constant.

### 3.2 Influence of treatment effect along the border

The treatment effect does not necessarily have to be constant along the border. In real-life applications, chances are that the treatment effect is variable along the border, especially if the border is long. This varying treatment effect is called spatial heterogeneity and may distort the estimate of the treatment effect if it is captured in one number.

We have experimented with four modifications of the treatment effect in simulations. The first simulation uses a baseline discontinuity of  $d = 8.0$ , constant at all boundary points. The second simulation tests the methods in absence of an effect: discontinuity is set to zero. In the third simulation, the discontinuity increases with one of the axes, which gives an equally spaced discontinuity with  $d = 0 \dots 10$  along the border. The last simulation tests the capability of the models to capture a sharp, temporal change to a high treatment effect, after which the true effect drops again. In this setting, the treatment effect is defined as Low - High - Low with corresponding effect values 4.0 - 8.0 - 4.0. The true effect will be 4.0, except for the region between 4 and 7 on the first

axis, where discontinuity is set to 8.0. In Figure 6, the true treatment effect along the boundary is shown, as well as the predictions of the discontinuity by the three different kernels. The white line shows the Bayesian Model Average line, which incorporates the posterior from every kernel to produce an average line, based on the plausibility of the fit to the data of that kernel. The BMA curve completely overlaps the fitted line from the Matérn kernel, indicating that the BMA tends to have strong beliefs in the correctness of the fit to the data by this kernel.

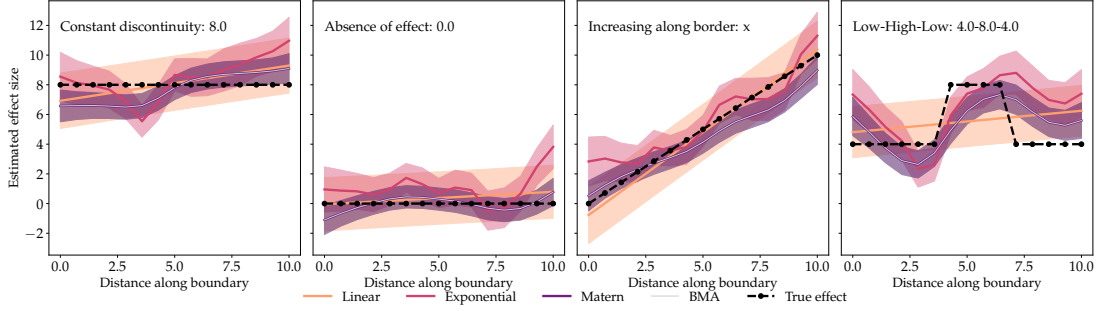


Figure 6: Plots of the true effect size and effect size estimation along the boundary per kernel for each of the four simulation specifications. In the top left corner of each plot, the treatment effect specification is shown.

As with the previous data modification, the methods are evaluated using the RMSE of the predictions along the boundary. Every simulation is run 100 times and the errors are averaged over these runs to minimize influences from the random factors. In Table 3, the averaged RMSE and the standard error is shown per measure for every kernel.

From these results, it appears that again the Matérn kernel consistently performs best. The Linear kernel provides the second lowest error and for every method the highest RMSE is generated by the exponential kernel, except for BNQD in the last simulation, in which the exponential kernel performs slightly better than the linear kernel. The exponential kernel, as is also obvious from Figure 6, yields a curve along the boundary with sharp turns and is not smoothly centered around the true effect size, causing a relatively large RMSE. The baseline effect (Simulation 1) is predicted best by the density-weighted LATE measure on a Matérn kernel, although all LATE measures show similar performance. The error produced by the 1D-RDD method, both in the standard and the bin-averaged case, is extremely large (4 to 10 times larger) compared to the errors produced by the other methods. In absence of an effect (Simulation 2), the BNQD BMA measure performs best. The error is, no matter the kernel, zero or very close to zero. As BMA explicitly accounts for the fact that the data may be drawn from a continuous model, and since a discontinuity is absent, this provides an accurate fit. Indeed, since a discontinuity is absent, the BMA tends to follow the continuous model, leading to a low RMSE. The performance of standard BNQD is slightly worse compared to the LATE measures. Again, it is reflected that the LATE measures perform slightly better in case of a constant treatment effect. Data that has a gradually increasing effect size along the border is studied in Simulation 3. From the errors produced it may be concluded that a continuously increasing effect is captured best by BNQD. In all three kernels, BNQD has the lowest average error. The BMA of BNQD produces an equally low error for the Linear and Matérn kernel, but its error is more than twice as

	Simulation 1			Simulation 2		
	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$
BNQD	0.718 (0.045)	1.279 (0.034)	0.649 (0.027)	0.722 (0.044)	1.245 (0.040)	0.680 (0.028)
BNQD BMA	0.718 (0.045)	1.279 (0.034)	0.649 (0.027)	<b>0.005</b> (0.005)	<b>0.000</b> (0.000)	<b>0.000</b> (0.000)
$\mathcal{W}_{UNIF}$	<b>0.515</b> (0.048)	0.722 (0.048)	0.348 (0.027)	0.534 (0.044)	0.612 (0.046)	0.359 (0.030)
$\mathcal{W}_{DENSITY}$	0.517 (0.044)	<b>0.650</b> (0.045)	<b>0.329</b> (0.026)	0.537 (0.044)	0.542 (0.042)	0.348 (0.028)
$\mathcal{W}_{INVVAR}$	<b>0.515</b> (0.048)	<b>0.650</b> (0.045)	0.337 (0.027)	0.534 (0.044)	0.547 (0.043)	0.354 (0.029)
$\mathcal{W}_{GEOM}$	<b>0.515</b> (0.048)	0.678 (0.047)	0.340 (0.027)	0.535 (0.044)	0.564 (0.044)	0.352 (0.029)
1D-RDD	3.362 (0.270)	4.216 (0.275)	3.430 (0.266)	0.822 (0.220)	1.017 (0.224)	1.048 (0.190)
1D-RDD avg	3.479 (0.270)	3.585 (0.285)	3.304 (0.265)	1.095 (0.224)	3.399 (0.280)	3.323 (0.262)
	Simulation 3			Simulation 4		
	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$	$K_{LIN}$	$K_{EXP}$	$K_{MAT}$
BNQD	<b>0.738</b> (0.043)	<b>1.249</b> (0.037)	<b>0.646</b> (0.025)	1.922 (0.017)	<b>1.494</b> (0.042)	<b>1.166</b> (0.025)
BNQD BMA	<b>0.738</b> (0.043)	2.442 (0.186)	<b>0.646</b> (0.025)	1.931 (0.021)	2.223 (0.140)	<b>1.166</b> (0.026)
$\mathcal{W}_{UNIF}$	3.161 (0.010)	3.177 (0.010)	3.110 (0.003)	<b>1.870</b> (0.014)	1.959 (0.021)	1.844 (0.010)
$\mathcal{W}_{DENSITY}$	3.173 (0.014)	3.177 (0.011)	3.124 (0.005)	1.871 (0.014)	1.981 (0.022)	1.859 (0.012)
$\mathcal{W}_{INVVAR}$	3.161 (0.010)	3.177 (0.010)	3.112 (0.004)	<b>1.870</b> (0.014)	1.964 (0.021)	1.852 (0.013)
$\mathcal{W}_{GEOM}$	3.170 (0.017)	3.172 (0.010)	3.115 (0.004)	<b>1.870</b> (0.014)	1.992 (0.023)	1.861 (0.013)
1D-RDD	5.007 (0.186)	5.831 (0.183)	5.532 (0.190)	3.873 (0.207)	4.903 (0.207)	4.236 (0.217)
1D-RDD avg	4.974 (0.184)	5.372 (0.239)	5.277 (0.234)	3.920 (0.202)	4.272 (0.251)	4.079 (0.246)

Table 3: Results of effect-size varying simulations. For every simulation, the average RMSE is shown for every tested method on each of the three kernels. The SEM is displayed within parentheses, the lowest error per simulation per kernel is shown in bold.

high in the Exponential kernel. As obvious from the third plot in Figure 6, this is due to the Exponential kernel shifting its estimation too drastically along the boundary, whereas the straight line from the Linear kernel and smooth Matérn fit follow the treatment effect more closely. The last simulation makes a sharp, temporary increase in treatment effect along the border. Because of this discontinuous shift in treatment effect, the models suddenly have to adapt to this fluctuation. Using an exponential or Matérn kernel, the lowest error is produced by BNQD, while the LATE measures

perform just slightly better than BNQD with a linear kernel. As the linear kernel is not able to make a quick change to a larger treatment effect, the model flows in between the two effect sizes and resembles an average (see Figure 6, rightmost plot), just as is produced by the LATE measures. The exponential and Matérn kernel give a better performance for BNQD, whereas the LATE measures perform equally well using either of the kernels. Apparently, neither kernel can make a discontinuous shift to a large treatment effect, causing the Matérn and exponential kernel to underestimate the effect when the treatment effect is low, and overestimate the effect when the effect is high, visible in Figure 6. Again, the distance to the border method produces large errors, showing a poor estimation for all three kernels. For all simulations, the standard error is largest in the distance-to-border errors, with a SEM of up to ten times the error of the LATE measures or BNQD. Again, the Matern kernel produces the lowest SEM in all simulations, indicating a low sample-to-sample variability of the RMSE.

There is not one outstanding method that performs best in all simulations. We may hence conclude that the treatment effect influences the performance of the methods in the GRD design. The LATE measures all produce values that are close to one another, while BNQD and the distance to border method may be much off in either positive or negative direction, compared to the LATE measures. The differences in average error for especially simulation 2 and 3 are rather extreme. This shows that the strength of BNQD is the ability to deal with changing effects, as is reflected in the results of Simulation 3 and 4. The LATE measures here all fail to cope with the changing discontinuity, although the errors are significantly lower than the error produced by the distance to border method. This method also cannot cope with spatial variation but has to base its estimation on a single intervention point, rather than on a set of boundary points. Thus, each category of discontinuity has a preferred method and this may be used as prior information when choosing a method and kernel. For example, if one suspects the data from having a heterogeneous treatment effect, one may use this information and advocate for the use of BNQD. The BMA of BNQD may also be used to find an estimate that will be less likely to overestimate the effect size if the treatment effect is relatively small compared to the noise level.

### 3.3 Influence of number of boundary points

The treatment effect size estimate is established using the estimates produced by the individual boundary evaluation points. Hence, a larger number of evaluation points to base the estimate on would presumably lead to a more accurate estimate of the effect. However, this would also imply a bigger number of computations to be made and therefore an increased computation time. It may not be necessary to evaluate the border at many points, as the treatment effect and data density are likely to be spatially correlated along the boundary.

In Figure 7, the performance of a GP with Matern kernel is shown against the number of boundary points used to compute the effect size, averaged over 100 runs. Here, a constant treatment effect of  $d = 10.0$  is assumed. The data are drawn from a uniform distribution with standard deviation  $\sigma = 1.0$ . This shows a sharp decrease in RMSE when the number of boundary points is low, but stops decreasing after the estimation is based on at least five evenly spaced boundary points. The variance on the other hand steadily decreases as the number of boundary points is increased, but

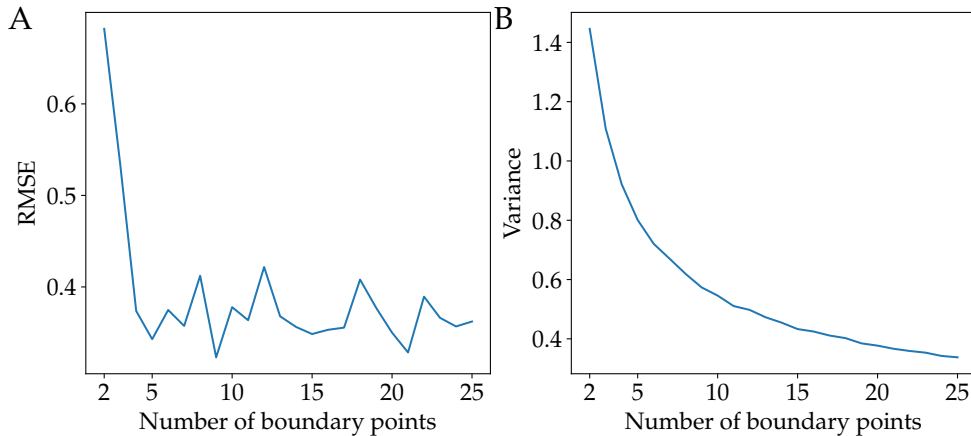


Figure 7: Model performance against the number of boundary points used to estimate the treatment effect, averaged over 100 runs. **A.** Root of Mean-Square Error from true discontinuity. **B.** Averaged variance per included number of boundary points.

after roughly 15 boundary evaluation points, the decrease is only very minimal. By increasing the number of points, the number of computations increases, leading to a trade-off between efficiency and accuracy. As the RMSE stops decreasing after 5 boundary points and the variance after about 15 boundary points, we take 15 boundary points as a good balance between preciseness and computational performance. We must note however, that in this case a constant treatment effect is assumed. If the treatment effect would vary along the border, presumably more boundary points would be needed to get an optimal accuracy-efficiency ratio, as the number of boundary points should at least be as big as the number of different treatment effects along the border, and all effects should be captured by at least one evaluation point.

## 4 Application

The relevance of the framework and its possible impact on society will be demonstrated using a real-life data set. The framework will be applied to a data set containing emission of ammonia in the Netherlands.

### 4.1 Emission of ammonia

One of the major concerns modern society is facing today is the preservation of our natural environment. Ammonia emissions into the environment negatively affect biodiversity and human health due to nitrogen pollution. In most of the ‘Natura2000’ natural reserve areas the deposition of nitrogen as a result of ammonia emissions exceeds the European norms. The agricultural sector is one of the main polluters when it comes to ammonia with 40% of the country’s emission and with the livestock farming industry as main source [26]. Central and regional governments have put several programs in place to reduce the ammonia emissions by enforcing strict regulations onto cattle sheds, by stimulating the use of modern technology and reducing the livestock as a whole.

Noord-Brabant and Limburg, two provinces in the south of the Netherlands, were appointed to become front runners in an all-embracing approach to improve the country's polluted air. In 2010, Brabant was home to over a third (34%) of all pig farms in the Netherlands, while Limburg accounted for almost 13% of all pig farms, amounting to nearly 8 million pigs in these two provinces [27]. Pig manure emits high amounts of ammonia, in traditional stables 3.0kg NH<sub>3</sub> per animal place per year, causing a tremendous local emission of ammonia and a high pressure on human health and environment. The Dutch national goal at the intensified policy, put into place mid 2015, was to reduce the emission of ammonia from pig farms by 47% for existing stables respectively 63% for new stables, leading to an emission of 1.6kg respectively 1.1kg NH<sub>3</sub> per animal place per year [28, 29]. This policy has already been implemented by Brabant and Limburg in 2010 [30, 31]. Moreover, the provincial policy of Brabant and Limburg prescribed an emission reduction of no less than 85% towards 0.45kg NH<sub>3</sub> per pig per year [28, 32]. A distinct difference in reduction levels in Brabant and Limburg versus the rest of the country by 2018 was to be expected. While Brabant actively pursued this goal, with a tightening of the policy in 2017 and forcing an emission reduction on all stables by 2020, Limburg took a more passive position towards reaching the target while its pig farming sector was to reach emission targets ultimately by the year 2030 [33]. This difference in attitude is expected to be reflected in their respective ammonia emission data.

We evaluate the policy effectiveness by creating a spatial-temporal RD design, where we take a look at the border between Brabant and Limburg. Before continuing the analysis, a number of assumptions requires validation, as discussed earlier. Near the border of the two provinces, there is no distinguishable difference in the number of pigs per square kilometer, which makes it reasonable to assume that any difference in emission is a result of the distinctive policy of the regions (thus validating the identification assumption). Without intervention, we would expect to see no difference in emission on either side of the border. This makes the identification assumption likely to hold. We also have no strong reasons to suspect that farmers have organized themselves near the border as regulations on both sides were equal to all farmers before the intervention in 2010 (validating the continuity assumption). The last assumption that needs validation in STRD designs is SUTVA (see Section 2.4). This assumption holds if the outcome of a unit does not depend on other units. In this application, one might be worried that the emissions of ammonia spread to nearby municipalities, violating the SUTVA assumption. However, most of the ammonia emission precipitates within its near surrounding [34], as can be seen in Figure 8A, in the case of Roermond (the one yellow circle in the east of Limburg) for example. Municipalities adjacent to Roermond all show a decrease in ammonia emissions, whereas the municipality of Roermond shows an increase of more than 20% over time. We may conclude that there is no effect of strong interaction between municipalities' outcomes. As neither of the three assumptions is violated we can proceed our case study.

A publicly available data set containing the emission of ammonia per year per municipality from the Dutch government is used for analysis of the ammonia concentration [35]. The ammonia emission for both provinces is retrieved for the years 2010 and 2018, which are the years in which the intensified policy was put into place, respectively the year that significant effects were to be expected. The data are pre-processed by divid-

ing the total emission per municipality by its geometric surface in square kilometers. Next, the emissions of each of the two years are reduced to one number by calculating the difference in emission between the years and using this value as outcome variable in the RD design. This reduces the temporal dimension from the spatial-temporal RD, which is the basic case of STRD, as explained in Section 2.4. The percentage decrease per municipality is calculated using the following formula:

$$d_i = \frac{E_i^{post} - E_i^{pre}}{E_i^{pre}} \quad i = 1 \dots N, \quad (9)$$

where  $N$  is the sum of the number of municipalities in Brabant and Limburg and  $E$  is the emission in the municipality, either before or after the intervention took place, respectively indicated with the superscript *pre* or *post*.

Several studies from both official and independent institutions have evaluated the effectiveness of the policy implemented in Brabant, e.g. the Dutch National Institute for Public Health and the Environment (RIVM), Connecting Agri & Food, and Bureau Polderoyen Compagnons [36–39]. These institutions came to the conclusion that the decrease in emission of ammonia is not as substantial as expected in advance or that an actual decrease is not even present. The reason for this absence of reduction is not known yet and is currently being investigated. One possibility is the effect of meteorological circumstances on the emission of ammonia. In the current setup of the GRD design, we cannot account for meteorological fluctuations. To identify a trend over multiple years rather than evaluating at two time points, one could incorporate a dimension of time and achieve a spatial-temporal RD design. As there are data points available only for a small set of years, the data is not extended with a temporal dimension and we stick with the percentage difference between 2010 and 2018.

The framework is applied to the data, where the provincial boundary is drawn as a line from south to north, with intermediate points to indicate corners. The boundary is depicted in Figure 8A as a black dashed line. Units to the west of this line are municipalities in Brabant, whereas units with a location to the east are located in Limburg. The framework creates two models per kernel and estimates the effect size and model probability for a polynomial with degree 1 (a linear kernel), exponential, and Matérn kernel. All three kernels fit the data separately for the continuous and discontinuous model. The predictions are evaluated at 30 ‘sentinel’ points along the boundary.

Effect of policy pursued in Noord-Brabant and Limburg			
<i>Kernel</i>	$p(m_0 D)$	$p(m_1 D)$	$\log BF_{10}$
Linear	0.002	0.998	6.24
Exponential	0.785	0.215	-1.29
Matérn ( $\nu = 3/2$ )	0.765	0.235	-1.18
BMA	0.687	0.313	-0.78

Table 4: Analysis of the ammonia emission policy. Displayed are the model posteriors for  $M_0$  and  $M_1$  and the log Bayes factors for each of the considered kernels and the Bayesian Model Average.

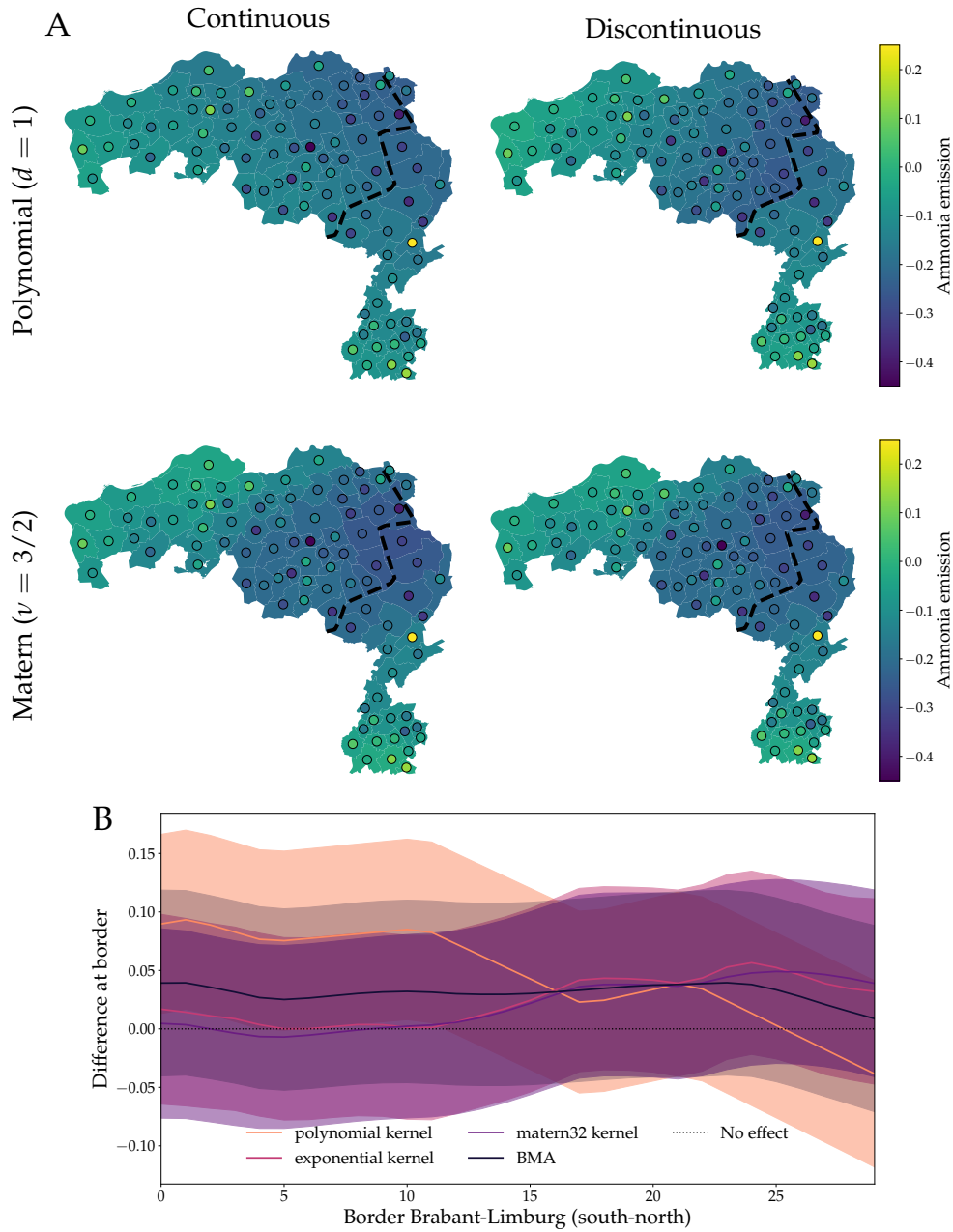


Figure 8: Discontinuity analysis along the province border, indicated by the dashed line. **A.** Circles indicate the observed percentage difference in ammonia emission between 2010 and 2018; municipalities are shaded according to the GP predictions. For conciseness of the figure, the fit of the exponential kernel is omitted. **B.** The predicted effect size  $p(d|D, M_1)$  along the province border for the three kernels and the Bayesian Model Average.

The results are shown in Table 4. As it appears, the linear kernel presents a Bayes Factor supporting the hypothesis that a discontinuity is present and thus that the Brabant policy is indeed more effective. The linear kernel is only able to fit a straight line along the boundary. However, since the border is not a straight line, it appears that the kernel presents a non-straight line along the boundary. The exponential and

Matern kernel on the other hand show a negative log Bayes Factor, indicating evidence *against* a difference in emission. As these kernels provide a better fit to the data than the linear kernel, the former two overshadow the Bayesian Model Average (BMA). In Figure 8B, the predictions are evaluated at every sentinel point and the differences in predictions of the discontinuous model along the border are displayed per kernel. As can be seen, the BMA line is a mixture of the three kernels, indicating that none of the kernels provides a flawless fit. Next to that, we observe that the linear kernel predicts a large difference along the south part of the border, while the absence of a difference and even a negative difference in the north indicates that the decrease in emission in Limburg was even larger than in Brabant. This can be fully accounted to the fit of the linear kernel, which is a flat surface on the data. Even though the linear kernel shows the presence of an effect, the exponential and Matern kernel lead the BMA. We use the BMA in order to arrive at the same conclusion as other institutions that have studied the effect, which is that there is no evidence that the policy pursued in Brabant has had a larger positive impact on the emission of ammonia. Rather, there is evidence showing that the policy in Brabant has led to an equal decrease of ammonia deposition as the less strict pursuing of the policy in Limburg.

## 5 Discussion

Worldwide, new governmental policies are implemented every day to create certain behaviour. The effectiveness of such policy can be hard to measure, as there is no direct way to randomly assign certain samples to the treatment group that receive the intervention and a control group. Under some assumptions, we are able to recreate the idea of random assignment and in this way evaluate the policy pursued. One application of such quasi-experimental design is geographical regression discontinuity (GRD) design, which splits the samples into a treatment and control group based on a boundary through geographical space. Units near the boundary are assumed to exhibit comparable characteristics and may therefore be compared. The difference in behaviour between units on both sides of the boundary is measurable as the difference between two fitted functions, extrapolated to the border. Where previous research focused on either reducing GRD to a 1D-RDD with distance to the border as assignment variable or capturing the treatment effect by averaging the local estimates, we here follow the approach taken by Hinne et al. (2020), called BNQD, and used Bayesian Model Averaging (BMA) to arrive at an effect curve along the border. This has the advantage that a treatment effect that varies along the border can be captured.

BNQD uses Gaussian Processes (GP) as a Bayesian non-parametric way for causal inference, whereas other researchers often have specified a less flexible polynomial function. GPs are described by their mean and covariance function, and this lays a, sometimes undesired, prior on the model. A wrongly chosen covariance function that provides a poor fit to the data can heavily influence the inferences made about the data. A partial solution to this problem is also presented in Hinne’s paper and involves BMA to automatically favour the kernel that fits the data best.

In the data simulation section, the advantages of GPs have become apparent. Simulations have shown that the estimates of the effect sizes in some cases were predicted best by BNQD, measured as the lowest error from the true effect size, and in other

simulation settings the LATE measures appeared more advantageous. BNQD can also explicitly take into account the probability that no effect is present, i.e. that the discontinuity is equal to zero, and this is done in the BNQD BMA step. In absence of effect, this null model overrules the alternative hypothesis and the error will be low. When only a small treatment effect is present, the BNQD BMA average likely underestimates the true effect size. A varying treatment effect along the border cannot be captured by the distance to border and LATE measure methods. Therefore, if the effect is likely to be heterogeneous along the border, it may be beneficial to use BNQD. Hence, it is important to think of the right method to use to produce the most accurate results.

GPs have also been used by Branson et al. [11] and in their follow-up paper specifically on GRD by Rischard et al [7]. The key difference between BNQD and the ‘sentinel point’ method used in these papers is the goal of the RD. Where we are here interested in finding evidence in favour of or against the null model or alternative model, their goal is to estimate the treatment effect and only test if the effect is significant. The effect found at each ‘sentinel point’ is weighed to produce one average estimation, from which no variations in treatment effect can be recovered. Other researchers (e.g., Black (1999), Lavy (2006), Lalive (2008), Keele (2015), and Salazar (2016)) have focused on modeling GRD with the help of linear regression or fitting a second or third order polynomial [3, 14–17]. In some of these papers the two dimensional space was reduced to a 1D-RDD, thereby losing the ability to capture spatial variation. As seen in the simulations in Section 3, the linear kernel often provides a relatively poor fit compared to the more complex, non-degenerate kernels, showing that linear regression might be a solution too simple to yield usable results.

Extending the spatially situated RD (GRD) design by adding a temporal dimension yields the novel spatial-temporal RD design, or STRD. This framework may be able to find the effects of an intervention based on both geographical location and time, allowing for a more powerful framework for causal inference. This new framework will have to deal with emergence of additional assumptions brought by the dimension of time and validation of the design becomes more complex. In Section 2.4, STRD is compared to related quasi-experimental design setups. The design is related to a Difference-in-Differences design in a spatial context. Assumptions identified for this design hence have to hold for STRD as well. The temporal dimension of STRD has similarities with Interrupted Time Series (ITS) analysis, although RD design focuses on the border between the control and intervention group, whereas ITS considers the bigger picture and uses all data points to establish an effect. For the more complex case of STRD, involving the addition of the dimension of time, a kernel has to be found that can fit both the geographical and temporal aspect of the data. Whether this is a feasible task is a subject for further research.

Application of BNQD in geographical settings is straightforward and has been shown in Section 4. The policy pursued by the provincial governments of Brabant and Limburg are evaluated using the percentage decrease of the emission of ammonia as variable per municipality and finding the difference of the extrapolations to the border. The Linear kernel found evidence in favour of a distinct difference between the two provinces, whereas the more complex exponential and Matérn ( $\nu = 3/2$ ) kernels did find evidence in favour of the null model. As the latter two kernels could more accurately adjust to the data, these dominate the BMA and lead to the conclusion

that the strict policy applied in Brabant did not lead to a larger decline of ammonia emissions. This example made use of the geographical location of municipalities and the province border as assignment threshold. The dimension of time is reduced by using the percentage difference between the emissions of 2010 and 2018, making this an application of the basic case of spatial-temporal RD. If more data would be present, sampled yearly or more often even, it could be beneficial to add the temporal dimension as additional dimension to the design to find a trend in the data and to minimize the influence of other factors, such as meteorological fluctuations that influence the amount of ammonia in air.

As a concluding remark we state that BNQD has its advantages and disadvantages over other methods in geographical contexts, particularly depending on the structure of and the discontinuity in the data. Nevertheless, it has been shown a valid approach in evaluating applied policies. This framework has the potential to improve policy analysis and may give policy makers a framework to base future decisions on.

## 6 References

- [1] Hinne, M., van Gerven, M. A., & Ambrogioni, L. (2019). Causal inference using Bayesian non-parametric quasi-experimental design.
- [2] Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2), 281-355.
- [3] Keele, L. J., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1), 127-155.
- [4] Geneletti, S., O’Keeffe, A. G., Sharples, L. D., Richardson, S., & Baio, G. (2015). Bayesian regression discontinuity designs: Incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in medicine*, 34(15), 2334-2352.
- [5] Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.
- [6] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.
- [7] Rischard, M., Branson, Z., Miratrix, L., & Bornn, L. (2020). Do School Districts Affect NYC House Prices? Identifying Border Differences Using a Bayesian Non-parametric Approach to Geographic Regression Discontinuity Designs. *Journal of the American Statistical Association*, (just-accepted), 1-35.
- [8] Donald B Rubin. (2005) Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.
- [9] Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201-209.

- [10] Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2), 200-215.
- [11] Branson, Z., Rischard, M., Bornn, L., & Miratrix, L. W. (2019). A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, 202, 14-30.
- [12] Duvenaud, D. (2014). The Kernel cookbook: Advice on covariance functions.
- [13] Paciorek, C., & Schervish, M. (2003). Nonstationary covariance functions for Gaussian process regression. *Advances in neural information processing systems*, 16, 273-280.
- [14] Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *The quarterly journal of economics*, 114(2), 577-599.
- [15] Lavy, V. (2006). From forced busing to free choice in public schools: quasi-experimental evidence of individual and general effects (No. w11969). National Bureau of Economic Research.
- [16] Lalive, R. (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of econometrics*, 142(2), 785-806.
- [17] Salazar, L., Maffioli, A., Aramburu, J., & Agurto Adrianzen, M. (2016). Estimating the impacts of a fruit fly eradication program in Peru: A geographical regression discontinuity approach (No. IDB-WP-677). IDB Working Paper Series.
- [18] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.
- [19] Delgado, M. S., & Florax, R. J. (2015). Difference-in-differences techniques for spatial data: Local autocorrelation and spatial interaction. *Economics Letters*, 137, 123-126.
- [20] Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3), 279-292.
- [21] Papay, J. P., Willett, J. B., & Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161(2), 203-207.
- [22] McDowall, D., McCleary, R., Meidinger, E. & Hay, R. (1980). Interrupted time series analysis. *Thousand Oaks*.
- [23] Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., & Reeves, D. (2015). Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *bmj*, 350, h2750.
- [24] Bernal, J. L., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology*, 46(1), 348-355.

- [25] Lee, D. K., In, J., & Lee, S. (2015). Standard deviation and standard error of the mean. *Korean journal of anesthesiology*, 68(3), 220.
- [26] Compendium voor de Leefomgeving (2018). Herkomst stikstofdepositie, 2018. Retrieved from <https://www.clo.nl/indicatoren/nl0507-herkomst-stikstofdepositie>.
- [27] Central Bureau for Statistics (2020). Statistics Netherlands: Landbouw; gewassen, dieren en grondgebruik naar regio [Data file]. Retrieved from <https://opendata.cbs.nl/#/CBS/nl/dataset/80780ned/table?dl=2C682>.
- [28] Ministerie van Infrastructuur en Milieu (2015). Besluit emissiearme huisvesting. Retrieved from <https://zoek.officielebekendmakingen.nl/stb-2015-266>.
- [29] Vermeij, I., Ellen, H., Bokma, S. (2017) Maatregelen ter reductie van ammoniakemissie in bestaande varkensstallen. Wageningen University & Research. Retrieved from <https://research.wur.nl/en/publications/maatregelen-ter-reductie-van-ammoniakemissie-in-bestaande-varkens>.
- [30] Provincie Noord-Brabant (2013). Regeling stikstof en Natura 2000. Retrieved from <http://decentrale.regelgeving.overheid.nl/cvdr/XHTMLoutput/Actueel/Noord-Brabant/CVDR276288.html>.
- [31] Zuidelijke Rekenkamer (2019). PAS (Programma Aanpak Stikstof) provincie Noord-Brabant. Retrieved from <https://zuidelijkerekenkamer.nl/pas-programma-aanpak-stikstof-provincie-noord-brabant/>
- [32] Provinciale Taskforce Stikstof (2020). Limburgs Aanvalsplan Stikstof: Op weg naar een nieuwe balans. Retrieved from <https://www.limburg.nl/thema/aanpak-stikstof/>
- [33] Berenschot en BügelHajema (2020). Beleidsevaluatie van het PAS en het wetstraject voorafgaand aan het PAS.
- [34] Rijksinstituut voor Volksgezondheid en Milieu. Retrieved from <https://www.rivm.nl/stikstof/vragen-en-antwoorden-over-stikstof-en-ammoniak>.
- [35] Emissieregistratie. (2018). ‘Emissiegegevens 1990-2018’ [Data file]. Retrieved from <http://www.emissieregistratie.nl/>.
- [36] Backus, G. & Sanden, A. v.d. (2017). Verwachte effecten aanpassen Verordening natuurbescherming en invoeren staldering op omvang en structuur veehouderij Noord-Brabant.
- [37] Meetnet Ammoniak Natuurgebieden. (2019). Meetresultaten Meetnet Ammoniak. Retrieved from [man.rivm.nl](http://man.rivm.nl).
- [38] Ullenbroeck, H. & Albers, K. (2017). Versnelling transitie veehouderij: effecten voor de natuur en het milieu.
- [39] Bleeker, A., Frumau A. & Hensen A. (2015). Datarapport Ammoniakmetingen 2007-2014.

## A Appendix

The following covariance functions have been used in this thesis:

$$\begin{aligned} k(x, x') &= \sigma^2 xx' + \gamma && \text{Linear} \\ k(x, x') &= \sigma^2 \exp\left(-\frac{x - x'}{\ell}\right) && \text{Exponential} \\ k(x, x') &= \sigma^2 \left(1 + \frac{\sqrt{3}(x - x')}{\ell}\right) \exp\left(-\frac{\sqrt{3}(x - x')}{\ell}\right) && \text{Matérn } (\nu = 3/2) \\ k(x, x') &= \sigma^2 \left(1 + \frac{(x - x')^2}{2\alpha\ell^2}\right)^{-\alpha}, && \text{Rational Quadratic} \end{aligned}$$

where  $\sigma^2$  is the variance of the GP,  $\gamma$  is the offset of the linear function,  $\ell$  is the length scale parameter and  $\alpha$  determines the relative weighting of large-scale and small-scale variations [12].