CHALMERS

Statistical Methods for Estimation of Paint Thickness and its Variance

Master's Thesis in Engineering Mathematics and Computational Science

HANNES MARLING

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2014

Abstract

IPS Virtual Paint at Fraunhofer Chalmers Centre (FCC) is a software for simulating electrostatic spray painting of objects. Simulating the painting process is extremely computationally demanding and hence very time consuming. It is therefore desirable to be able to obtain results based on as few number of simulations as possible. This, and the random behavior of the simulations, give rise to randomness in the estimated paint thickness that is not easily quantified by means of ordinary methods. The purpose of this thesis is to further develop methods for estimating the resulting paint thickness and its variance.

Different varieties of models based on nonparametric kernel density estimation for estimating of paint thickness were evaluated. This was done using synthetic data, along with data generated via IPS Virtual Paint. Variations of anisotropic kernel estimations for reducing bias in the estimates were also investigated. Also, methods for enhancing existing methods when performing paint thickness estimations along the edges of an object were developed. Both the anisotropic methods together with the edge compensation algorithms showed to improve the quality of the estimates.

Previous work at FCC have used regression models for estimating the variance of the estimated paint thickness. This method is however not applicable for the entire object. In this report an alternative method, based on bootstrap, for estimating the variance is analyzed. The results show that bootstrap performs well for the different scenarios investigated, with results consistent with regression models and more exact estimates produced through numerous painting simulations.

Acknowledgments

I would like to express my deep gratitude to my supervisors Fredrik Ekstedt and Björn Andersson at FCC, for guidance, inspiration and critiques of this thesis throughout the process. I would also like to thank everyone at FCC for letting me do my thesis here and also providing with a friendly and welcoming atmosphere and with the necessary tools. Thanks to my friend David Eriksson at FCC for giving me advices and help when I got stuck during the work. I would also like to thank my examiner Prof. Patrik Albin at Chalmers University of Technology for reading my thesis and giving me valuable feedback.

Finally, I wish to thank my parents, Rolf Marling and Annika Svahn, for their support and encouragement throughout my time studying at Chalmers.

This work has been conducted within the Virtual Paint Shop project with the partners Volvo Cars, Scania, AB Volvo, Swerea IVF and BT Products. It was supported in part by the Swedish Governmental Agency for Innovation Systems, VINNOVA, through the FFI Sustainable Production Technology Program and in part by the Sustatinable Production Initiative and the Production Area of Advance at Chalmers. The support is gratefully acknowledged.

Hannes Marling, Göteborg June 2, 2014

Contents

Notation

Here follows a brief list of some of the common notations used throughout the report to reduce the risk of possible confusion:

- $\boldsymbol{x}$ - (x, y) or (x, y, z) coordinates on a 2 respective 3 dimensional surface.
- S - The surface of the object being painted.
- $\boldsymbol{X}$ - (x, y) or (x, y, z) coordinates for the impact of a droplet on S.
- V - The volume of a paint droplet.
- γ - *Cloud factor*, ratio between the real and simulated number of droplets, $\gamma \geq 1$.
- h - *Bandwidth* parameter in the kernel estimation model.
- λ - *Smoothing* parameter used in thickness estimation.
- $\tau(\boldsymbol{x})$ - The paint intensity at $\boldsymbol{x}$.
- $\widehat{T}(\boldsymbol{x}; \gamma, h)$ - Estimated paint thickness at $\boldsymbol{x}$ for the $h-$model.
- $\widehat{T}(\boldsymbol{x}; \gamma, \lambda)$ - Estimated paint thickness at $\boldsymbol{x}$ for the $\lambda-$model.
- σ^2 - The variance of $\widehat{T}(\boldsymbol{x}; \gamma, .)$.
- $\widehat{\sigma}$ - Estimate of the standard error σ.
- c_v - Coefficient of variation, normalized standard deviation.
- η - Relative error of the estimate $\widehat{\sigma}$ using the simulated values as true values of σ.
- η_s - Relative error taken with sign.

1

Introduction

The software IPS (Industrial Path Solutions) Virtual Paint, at Fraunhofer Chalmers Research Centre (FCC), models among other things electrostatic spray painting of cars and other objects [1, 8]. This is a complex process which involves advanced physical modeling and is computationally demanding. Electrostatic spray painting uses charged droplets injected at high speed from a nozzle towards a grounded body. The reason for charging the droplets is to get a more uniform coating on the body and to increase the transfer efficiency [18]. Each simulation results in a large number of droplet impacts on the body. Randomness is incorporated in the simulation, both for the volumes of the droplets and their resulting trajectories from the nozzle to the body through the air. Therefore the result of a simulation will differ from previous ones. Using mathematical models, the resulting paint thickness is then estimated from these impacts.

To increase the computational efficiency of the simulation, which in full scale would have to handle an enormous number of droplets, a smaller amount of droplets are simulated; an actual spray painting of a car uses about 10^{14} droplets. Using less droplets introduce more randomness to the modeling which will influence the uncertainty of the estimated paint thickness. It is therefore of great importance to be able to measure the effect of this uncertainty, and hence the resulting precision of the estimated thickness. For practical aspects, methods for estimating the precision by only using data generated from only one simulation, is desired. Using multiple simulations to obtain the amount of data required to produce accurate estimates would simply consume too much time and computational effort when the software is used in an industrial project.

1.1 Earlier work

Modeling and simulation of rotary bell spray atomizers in automotive paint shops is given in the doctoral dissertation by Andersson [1]. It encapsulates much of the existing research in the area and is a good starting-point for the interested reader.

In the Bachelor's thesis by Andersson and Johansson [2], the use of statistical models for describing the paint thickness on a surface is developed. The modeling of the droplets on the surface is done by the use of Poisson processes, with each of the droplets having log-normally distributed volumes. Our approach when generating synthetic data is highly influenced by their work.

The article by Tafuri et al. [18] initiates the use of kernel density estimation methods for performing paint thickness estimations. This showed to give better estimates than previous methods using histogram based methods. An anisotropic method for improving the estimation was also evaluated and did successfully reduce bias of the estimated paint thickness near the edges of the investigated surfaces. The idea behind the use of the anisotropic method is mainly based on the work given in Schjøth et al. [16], which uses similar methods for diffusion based photon mapping. Chapter 3 of this thesis incorporates many of the ideas given in these articles.

In a former Master's thesis by Isaksson [8] at FCC, methods for estimating the variance of the paint thickness using local regression models were developed. These methods showed to work well for estimating the variance at inner points where the surface and the paint thickness did not vary too much. The models did however not take into account general cases, for instance how boundary effects will influence the paint thickness estimation and its precision, or alternative methods for points where the assumptions for the regression are not reasonable. In Chapter 4 an alternative method to regression for variance estimation will be developed and evaluated by comparing with regression based methods.

1.2 Problem formulation and purpose

The purpose of this thesis is to further develop and evaluate methods for estimating the paint thickness and its variance along surfaces painted through electrostatic spray painting. We investigate the difference performance between two different models and how these depend on parameters used. We are also interested in trying to incorporate edge and boundary effects in the estimations to achieve better estimations of the paint near boundaries than those existing today. Existing methods for estimating the variance will be further investigated and compared to an alternative method based on bootstrap.

1.3 Limitations

This thesis does not take into account the complex painting process, but only the resulting simulated data. It uses both data which has been generated through `MATLAB` by statistical methods, and data generated by the software IPS.

The algorithms were only developed and tested for two dimensional surfaces, which is a simplification of the actual cases. When performing thickness estimations for three dimensional surfaces, no corrections for the arising complications were taken into account. The reason for these simplifications is that it would have been taking too much time and effort compared to the possible gain. Our judgment is that the results and conclusions drawn from two dimensional data is valid also in three dimensions.

1.4 Mathematical description of the problem

Let S be a surface in $\mathcal{R}^3$ and let $\boldsymbol{X}_d \in S$ and $V_d \in \mathcal{R}_+, d = 1, ..., n$ be a set of n impact locations of paint droplets on the surface and their corresponding volumes. The impact locations can be characterized by an inhomogeneous spatial Poisson point process on S with an intensity function $\nu : S \to \mathcal{R}$ [2].

Each droplet d is associated with a volume $V_d \in \mathcal{R}_+$, distributed according to a log normal distribution with an upper limit. Combining both the spatial process for the impact locations and the volume process, this can be seen as a Marked Poisson point process

$$X = \{(d,V_d) : d \in S\},$$

with points in S and mark space $\mathcal{R}_+$ [17].

Let $A \subset S$ be a part of the surface and denote by

$$V_A = \left\{ \sum_d V_d \ ; \ X_d \in A \right\}$$

the total volume of all droplets with impacts on A. By assuming a *paint thickness intensity* $\tau : S \to \mathcal{R}$ we can write

$$\mathbf{E}(V_A) = \int_A \tau \, \mathrm{d}S,$$

i.e. that the expected total paint volume of A can be obtained by integrating the intensity over A.

Using physical models of the spread of a droplet's impact on a surface and the data sets $\boldsymbol{X}_d$ and V_d, the resulting paint thickness on S, a random variable denoted by $T(\boldsymbol{x})$, would be computable. Our goal is to estimate $\mathbf{E}[T(\boldsymbol{x})]$, which will be a smoothed version of $\tau(\boldsymbol{x})$. Ideally, this would however require physical models that will not be used. Subsequently, when referring to an estimate of the thickness, what we mean is an estimate of $\mathbf{E}[T(\boldsymbol{x})]$.

As the total number of droplets needed when painting an object is enormous, only a fraction of the actual droplets can be simulated in a reasonable amount of time. The fraction between the real and simulated number of droplets is called *cloud factor* and denoted by γ throughout this thesis. This will introduce a new source of randomness depending also on γ; an estimate of the paint thickness at $\boldsymbol{x}$ will throughout be denoted by $\widehat{T}(\boldsymbol{x},\gamma)$.

We are also interested in specifying the precision in the estimates of $\widehat{T}(\boldsymbol{x},\gamma)$ *i.e.* in estimating Var $\widehat{T}(\boldsymbol{x},\gamma)$. Subsequently the variance of $\widehat{T}(\boldsymbol{x},\gamma)$ will be denoted σ^2 and an estimate of σ (*i.e.* the standard error) will be denoted by $\widehat{\sigma}$.

2

Test scenarios

This chapter will briefly introduce the different test scenarios that will be under consideration during the report.

2.1 Synthetic data

`MATLAB` will be used to generate synthetic data in two dimensions. By defining an intensity $\nu(\boldsymbol{x})$, realizations of a painting process can be mimicked by sampling spatially distributed impact locations according to a spatial Poisson point process with intensity $\nu(\boldsymbol{x})$ on the surface S. This can be done for arbitrary intensities $\nu(\boldsymbol{x})$ by using acceptance-rejection sampling, see Appendix A.1. To each impact a volume V is assigned randomly and independently, where V is a log normally distributed random variable but with an upper bound. For simplicity, V is assumed to be independent of $\boldsymbol{x}$. Hence we can define a paint thickness intensity $\tau : S \to \mathcal{R}$ as

$$\tau(\boldsymbol{x}) = \mathbf{E}[V]\nu(\boldsymbol{x}).$$

The cloud factor γ will be used to scale the volumes of each droplet when performing estimates, hence a droplet with volume V will contribute with γV volume of paint.

The number of droplets needed, denoted by n below, is calculated by first finding the total volume of paint needed through

$$V_{tot} = \int_S \tau \, \mathrm{d}S.$$

Now the number of droplets can be calculated as

$$n = \frac{V_{tot}}{\gamma \mathbf{E}[V]}.$$

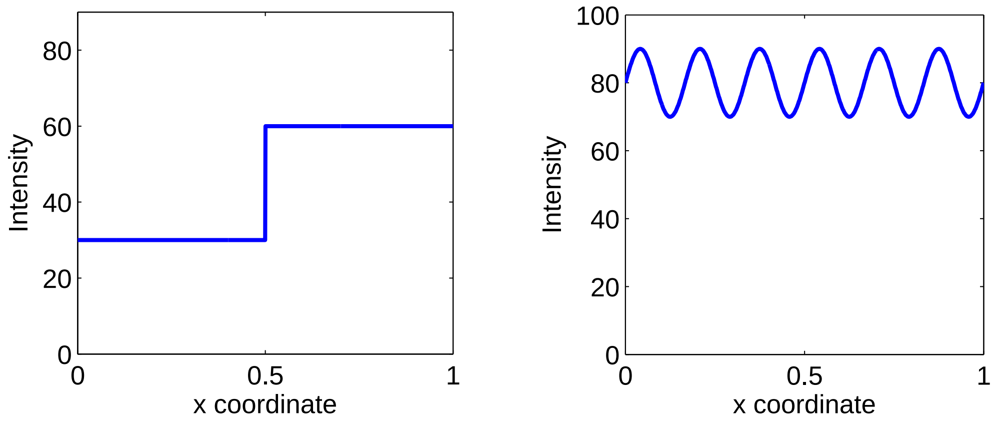
The surface S used in the synthetic scenarios is chosen to be a square with sides 1 m. Two different choices of the intensity $\tau(\boldsymbol{x})$ will be used. In the first scenario (denoted by **S-I** for further reference), the intensity is

$$\tau(x,y) = \begin{cases} 30, 0 \leq x < 1/2 \\ 60, 1/2 \leq x < 1 \end{cases}$$

and in the second scenario (denoted **S-II**) the used intensity is

$$\tau(x,y) = 80 + 10\sin(2\pi k x),$$

where k is a parameter related to the wavelength. In Figure 2.1(a) and Figure 2.1(b) the intensity functions can be seen as a function of x only.



(a) The intensity function for the **S-I** scenario

(b) The intensity function for the **S-II** scenario, using $k = 6$

Figure 2.1: Intensities for the two different synthetic test scenarios

Using the **S-I** scenario will give information on how the investigated methods perform when there is an abrupt change in the thickness along the surface, as is the case close to the step in Figure 2.1(a). On the other hand, apart from the step, the thickness is flat (but the randomness will of course not make it perfectly flat). When performing actual spray painting of an object, this is desirable, so hence it is also interesting to see how the methods can handle this case. In the **S-II** scenario the wave-shaped intensity is meant to resemble the variations in the thickness along the surface arising naturally from the nozzle profile and the path of motion of the painting robot. In contrast with the previous test scenario, the intensity varies much more rapidly, but with a lower order of magnitude. Hence, it is to expect that it will require more of the algorithms to detect the variations in this scenario.

2.2 Data generated by IPS

IPS was used to generate more realistic data. In the first scenario (**IPS-I** for further reference) a square with the same dimensions as in the synthetic case were used. Figure 2.2 shows the simulation of the painting process for this case in IPS when applying one stroke of paint from the painting robot onto the surface. There is a second plate placed above the plate orthogonal to the direction of the path of the painting robot. This will however not be taken extra care of; the surface of this extra plate will be treated as if it would be the surface of the square (since it will not matter for anything done in this thesis).

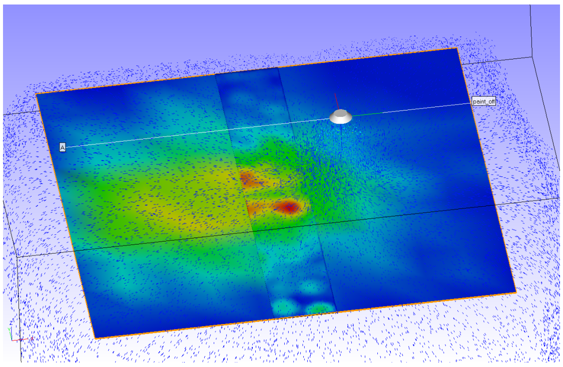


Figure 2.2: IPS simulating the painting process in the **IPS-I** scenario

The major difference between the data obtained by IPS and the synthetic data is that the underlying stochastic process which generates the impact locations no longer have a simple appearance in form of an intensity τ. The random behavior in the paint originates from randomness in velocities, directions and volumes for the paint when exited from the nozzle. To calculate the trajectories through the air highly nonlinear chaotic PDE's has then be solved numerically. This together with the electromagnetic fields makes it impossible to try to convert the randomness in the painting process to an intensity living on S. However, conceptually there is an intensity $\tau(\boldsymbol{x}), \boldsymbol{x} \in S$, but it is not available. Other differences is that the volumes are no longer independent of the impact locations $\boldsymbol{x}$. For more information about the modeling of the painting process, see [1].

The same plate were also painted using five strokes of the painting robot, yielding a more even coating of paint. This scenario will be denoted **S-II** for further reference.

Simulating the painting of an actual object

Finally, IPS was used to simulate the painting process of a Volvo V60. In Figure 2.3 the ongoing simulated can be seen. For further reference, this will be denoted the **IPS-III** scenario. Although scaling down the simulation by using a fairly large cloud factor to reduce the number of simulated droplets needed, the simulation required a lot of time; it takes about 40 hours on a modern computer utilizing the graphics card for computationally intensive tasks to complete such a simulation. This makes it obvious that it is not feasible to carry out several simulations to perform estimates of the variance in the resulting paint thickness between simulations. Other methods are needed to complete this task.

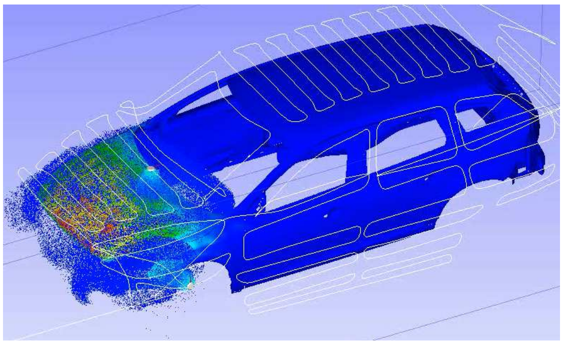


Figure 2.3: IPS simulating the painting process in the **IPS-III** scenario (geometry courtesy of Volvo Car Corporation)

3

Paint thickness estimation

This chapter will present models for estimating the paint thickness using kernel density methods. The different test scenarios presented above will be used. First the kernel density estimation method will be adjusted so that they can be used for estimating paint thickness, see Appendix A.2 for more information regarding kernel density estimation. Two conceptually different models for the estimation will be investigated using the synthetic data. Methods for estimating the paint thickness at more problematic points on the surface (to be defined later) will then be investigated and evaluated.

3.1 Estimation of the paint thickness using kernel estimates

The kernel density estimate method from Appendix A.2 produces a smooth estimate of a probability density function from a set of realizations of a random variable. Analogous methods also work for estimating the paint thickness from a set of droplet impacts on a surface by means of the estimated density function. Given a set of impact locations and volumes of droplets, we want to convert these to something which can be interpreted as a thickness. This is indeed very similar to the usual problem using kernel density estimation. The estimates will however not be normalized to integrate to 1 anymore, which is the case for a probability density function, but to integrate to the total volume of paint on the surface. To calculate the paint thickness at a point $\boldsymbol{x}$ on the surface, nearby impacts within a distance h, are located using a $kd-$tree [9]. For each nearby impact, a kernel function is used to calculate its contribution to the thickness at $\boldsymbol{x}$. Since a kernel function is such that it integrates to 1 (by construction), we have to weight the contribution with the volume V of the droplet in order to get the correct contribution of paint on the surface. By summing up, this would give the desired thickness if γ would be 1, but as this is not the case the estimate have to be weighted also with γ to correct for the smaller amount of impacts. Equation 3.1 shows the estimated thickness at $\boldsymbol{x}$

as a function of γ and the bandwidth h for the kernel function $\mathcal{K}$ using this method, where the $\boldsymbol{X}_i'$s are the coordinates of droplet i's impact on S. It is important to note that the interpretation of the bandwidth is not the physical spread of the droplet, but the neighborhood of $\boldsymbol{X}_i$ affected by the impact (*i.e.* the subset of S within a distance h from $\boldsymbol{X}_i$), which due to the large cloud factor is typically of much larger magnitude than the actual physical spread of the droplet.

$$\widehat{T}(\boldsymbol{x};\gamma,h) = \gamma \sum_{i=1}^{n} \frac{V_i}{h^2} \mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right). \tag{3.1}$$

In the model given in Equation 3.1 a large cloud factor and large droplets results in peaks in the estimated thickness, since the greater volumes and cloud factors are incorporated only as an increase in the amplitude of its contribution to the thickness on the surface. This may be undesirable since the log-normal distribution will produce some relatively large droplets, and hence this method could lead to unstable estimates.

Another more adaptive kernel estimation model for the paint thickness is to let the size of each droplet and the cloud factor determine also the spread of the droplet as it hits the surface (where the spread again should not be interpreted as the physical spread but the domain of S which will be influenced by the droplet at $\boldsymbol{X}_i$). A smoothing factor, λ, will also be included for tuning the model, which can be seen in Equation 3.2, with r_i being the radius of droplet i. Here larger volumes are partly included as a wider impact spread, proportional to each droplet's radius.

$$\widehat{T}(\boldsymbol{x};\gamma,\lambda) = \gamma^{1/3} \sum_{i=1}^{n} \frac{r_i}{\lambda^2} \mathcal{K}\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{\lambda r_i \gamma^{1/3}}\right). \tag{3.2}$$

The kernel function that will be used in this thesis is the two dimensional radial symmetric Epanechnikov kernel $\mathcal{K}$, constructed from the one dimensional Epanechnikov kernel K through $\mathcal{K}(\boldsymbol{x}) = \alpha K(\|\boldsymbol{x}\|)$. In two dimensions $\alpha = \pi/2$, see Appendix A.2 for more details. The function $K(x)$ can be seen in Figure 3.1.

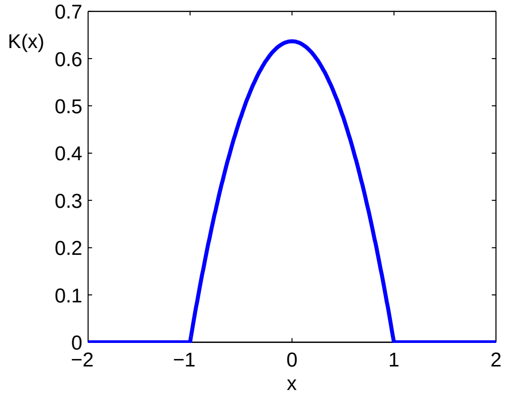
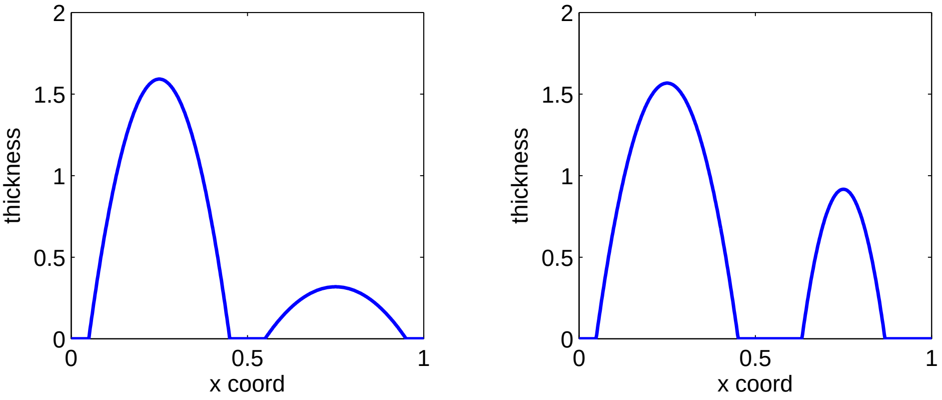


Figure 3.1: The one dimensional Epanechnikov kernel, $K(x)$

In Figure 3.2 the different behavior of the two models can be seen when estimating the resulting paint thickness from two droplets on a surface, where the larger droplet is 5 times as large as the smaller. The figure shows a cut along the x axis when placing the two droplets aligned along the y axis. In the $\lambda-$model, the smaller droplet has a narrower spread than the larger, compared to the $h-$model where the drops have equal spread. However, in the $h-$model, the amplitude of the smaller droplet is 5 times smaller than for the larger, in contrast to the $\lambda-$ model, where the difference in amplitude is smaller. Note that the total volume for each droplet is the same regardless of which of the two models that are used.



(a) $h-$model (b) $\lambda-$model

Figure 3.2: The different behavior of the $h-$ and $\lambda-$model

3.2 The two different models

To investigate the performance of the two different models presented above for estimating the paint thickness the two test scenarios using synthetic data were used. To perform estimations, a reasonable value of the bandwidth and the smoothing factor has to be set. If they are set too low, the resulting estimate will be very noisy; choosing them too large on the other hand will increase the error due to bias. Figure 3.3 shows the estimates using a too large (blue) and a too small (red) value of h respectively in the $h-$model for the **S-II** scenario, based on the average of 3 horizontal lines along the x axis, together with $\tau(\boldsymbol{x})$ for $k = 2$. The estimate using the larger value suffers from bias, while the estimate using the smaller value suffers from high variance.

Clearly, choosing h (or λ) correctly is crucial for the quality of the estimation, and for a given cloud factor, we want to be able to determine how to optimally choose h and λ respectively in the two scenarios. By optimal here we will use the mean integrated squared error, MISE, as measure of the (global) error [7]. Hence, the values of h and λ yielding the lowest value of the MISE for a specific cloud factor γ will be considered as optimal.

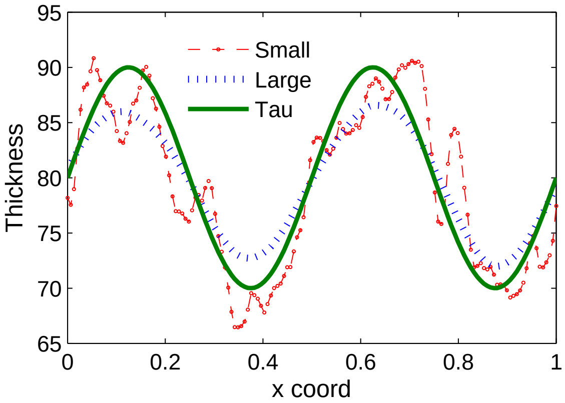


Figure 3.3: Estimates using the h-model with a large and small value of h respectively, together with $\tau(\boldsymbol{x})$ for the **S-II** scenario

To optimize the two models for estimating the paint thickness for inner points with respect to h and λ respectively, droplets were generated on a larger surface containing the surface S, to be used in Equations 3.1 and 3.2 respectively. This removes the problem of the lack of neighboring points when estimating the thickness close to boundaries, therefore the entire surface could be considered to solely consist of inner points when calculating the estimates.

For a wide range of cloud factors, the optimal values of h and λ will be calculated. This will yield optimal values of the MISE, denoted MISE_h and MISE_λ respectively in the following sections, for each cloud factor chosen. First the squared error, SE, is calculated as

$$\text{SE} = \left(\widehat{T}(\boldsymbol{x};\gamma,.) - r(\boldsymbol{x})\right)^2, \tag{3.3}$$

where r is a reference thickness. Ideally we would want to use $\mathbf{E}[T(\boldsymbol{x})]$ as the reference thickness; this is however not possible since it is unknown. Therefore other quantities have to be used. Integrating the SE over S, the integrated squared error, ISE, is obtained

$$\text{ISE} = \int_S \text{SE}\, \mathrm{d}S.$$

Because of the randomness, ISE is a random variable which connects to the MISE as

$$\text{MISE} = \mathbf{E}[\text{ISE}].$$

When optimizing for the synthetic data the intensity τ will be used as the reference thickness r. Later when performing similar optimizations for the data generated by IPS, a reference thickness will instead be constructed by averaging over many simulations using small values of the bandwidth to reduce the bias. It is not obvious which of the above choices of r that is most reasonable. Comparing with the intensity has the advantage that bias will contribute to the error, which it should. However, it is not reasonable to expect the thickness to behave like the intensity. Comparing with a reference thickness obtained by averaging over many simulations will on the other hand partly correct for the bias. What makes it hard when choosing method is that we do not know the actual thickness; based on the available information this is the best we can do.

The $h-$model

The integrated squared error, ISE, was calculated for different values of the cloud factor γ when varying h in Equation 3.1 by first subtracting $\tau(\boldsymbol{x})$ from $\widehat{T}(\boldsymbol{x};\gamma,h)$ in Equation 3.3, yielding the squared error SE, and then computing a numerical approximation of the integral of SE over S. By repeating the procedure we get several realizations of the ISE and averaging over them yields an estimate (by means of the law of large numbers, [15]) of the MISE as

$$\text{MISE} \approx \frac{\text{ISE}_1 + \text{ISE}_2 + ... + \text{ISE}_n}{n}.$$

Cubic spline interpolations were used to get the approximate MISE as a continuous function of h which could then be minimized in order to obtain the optimal values [12]. The trade-off between variance and bias is evident. In Table 3.1 the optimal values of h together with the corresponding $\sqrt{\text{MISE}_h/\text{MISE}_\infty}$ for varying values of γ can be seen for the case with the thickness distributed according to the **S-I** scenario. MISE_∞ is the theoretical MISE we would obtain by using the theoretical average thickness in Equation 3.3 obtained by averaging the intensity τ over S, instead of using the estimated thickness. Hence, MISE_∞ is calculated by letting

$$\widehat{T}(\boldsymbol{x};\gamma,h) = \int_S \tau(\boldsymbol{x})\,\mathrm{d}S \Bigg/ \int_S \mathrm{d}S$$

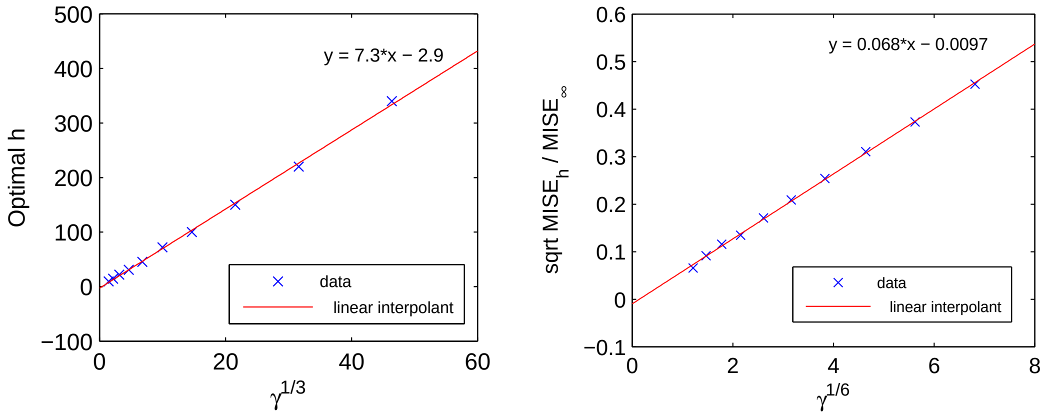
in Equation 3.3. $\sqrt{\text{MISE}_h/\text{MISE}_\infty}$ can hence be interpreted as a measure of how much of the variation in the paint that the model fails to capture. In Table 3.2 the corresponding table for the **S-II** scenario can be seen for two different values of k.

In Figure 3.4(a) the optimal values of h as a function of $\gamma^{1/3}$ for the step function scenario can be seen together with a linear interpolation, which seems to be a good way of describing their relationship. We can also get a linear relationship between $\sqrt{\text{MISE}_h/\text{MISE}_\infty}$ and $\gamma^{1/6}$ to be seen in Figure 3.4(b).

In Figure 3.5 similar plots for the **S-II** scenario can be seen for $k = 3$. Now the relationship between $\sqrt{\text{MISE}_h/\text{MISE}_\infty}$ and $\gamma^{1/3}$ seems to be linear.

Table 3.1: Values of h that minimizes the MISE for different values of γ for the **S-I** scenario

γ	h (mm)	$\sqrt{\mathrm{MISE}_h/\mathrm{MISE}_\infty}$
10^5	340	45.3%
$10^{4.5}$	220	37.3%
10^4	150	31.0%
$10^{3.5}$	100	25.4%
10^3	72	20.9%
$10^{2.5}$	46	17.1%
10^2	31	13.5%
$10^{1.5}$	22	11.6%
10^1	15	9.14%
$10^{0.5}$	9.4	6.55%



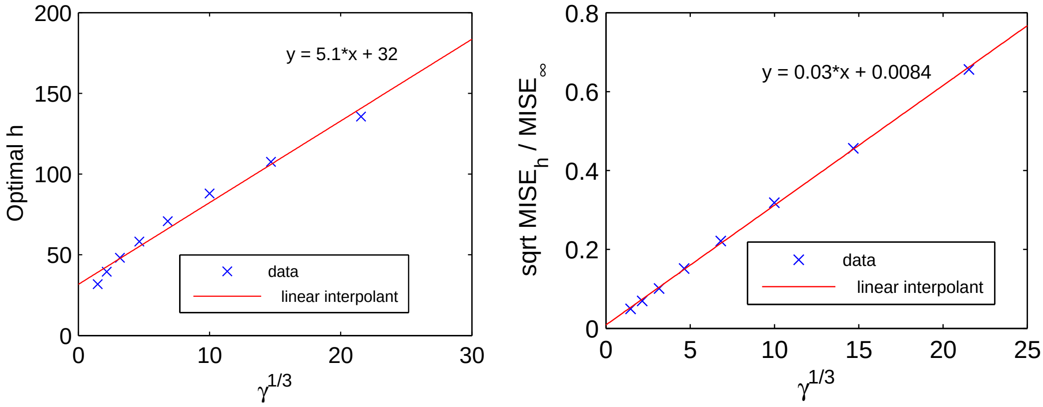
(a) Optimal values of h in mm as a function of $\gamma^{1/3}$ together with a linear interpolant

(b) $\sqrt{\mathrm{MISE}_h/\mathrm{MISE}_\infty}$ as a function of $\gamma^{1/6}$ together with a linear interpolant

Figure 3.4: Results for the $h-$model in the **S-I** scenario

Table 3.2: Values of h that minimizes the MISE for different values of γ for the **S-II** scenario for different values of k

	γ	h (mm)	$\sqrt{\text{MISE}_h/\text{MISE}_\infty}$
k=3	10^4	140	65.6%
	$10^{3.5}$	110	45.6%
	10^3	88	31.8%
	$10^{2.5}$	71	22.1%
	10^2	58	15.2%
	$10^{1.5}$	48	10.0%
	10^1	40	6.95%
	$10^{0.5}$	32	4.92%
k=10	10^3	42	152%
	$10^{2.5}$	33	106%
	10^2	27	73.2%
	$10^{1.5}$	22	50.5%
	10^1	18	34.6%



(a) Optimal values of h in mm as a function of $\gamma^{1/3}$ together with a linear interpolant

(b) $\sqrt{\text{MISE}_h/\text{MISE}_\infty}$ as a function of $\gamma^{1/3}$ together with a linear interpolant

Figure 3.5: Results for the $h-$model in the **S-II** scenario for $k = 3$

The λ-model

Similar simulations as above were also made for the $\lambda-$model using Equation 3.2. In Table 3.3 the optimal values obtained for λ together with the corresponding values of $\sqrt{\text{MISE}_\lambda/\text{MISE}_\infty}$ can be seen in the case with the paint varying as a step function and in Table 3.4 the corresponding table for the sinusoidal paint distribution can be seen.

Table 3.3: Values of λ that minimizes the MISE for different values of γ for the **S-I** scenario

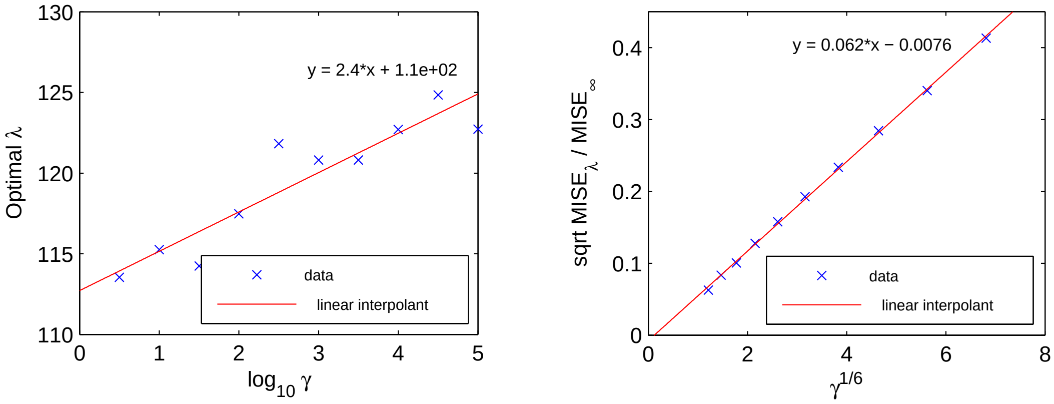
	γ	λ	$\sqrt{\text{MISE}_\lambda/\text{MISE}_\infty}$
	10^5	123	41.3%
	$10^{4.5}$	125	34.0%
	10^4	123	28.4%
	$10^{3.5}$	121	23.3%
	10^3	121	19.3%
	$10^{2.5}$	122	15.8%
	10^2	117	12.8%
	$10^{1.5}$	114	10.0%
	10^1	115	8.34%
	$10^{0.5}$	114	6.28%

Table 3.4: Values of λ that minimizes the MISE for different values of γ for the **S-I** scenario with $k = 3$

	γ	λ	$\sqrt{\text{MISE}_\lambda/\text{MISE}_\infty}$
k=3	10^4	113	60.3%
	$10^{3.5}$	129	41.9%
	10^3	152	29.4%
	$10^{2.5}$	181	20.0%
	10^2	215	14.0%
	$10^{1.5}$	258	9.46%
	10^1	308	6.44%
	$10^{0.5}$	374	4.57%

Figure 3.6(a) shows the optimal values of λ as a function of $\log_{10}\gamma$ for the **S-I** scenario. Optimal values of λ seems to not depend much on the values of γ. There is a slightly positive correlation between larger γ and larger λ, but over all a value of about

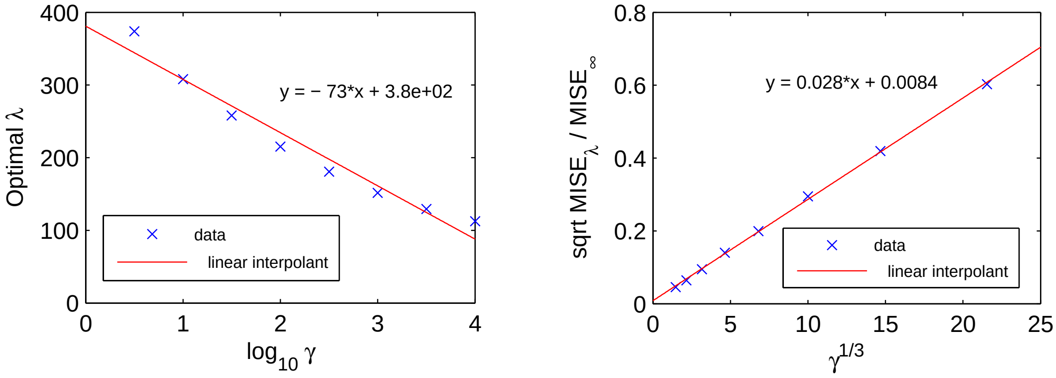
$\lambda \approx 120$ could be used for all cloud factors in the range investigated. Figure 3.6(b) shows $\sqrt{\text{MISE}_\lambda/\text{MISE}_\infty}$ against $\gamma^{1/6}$, and here the relationship is linear. In Figure 3.7(a) optimal values of λ against $\log_{10}\gamma$ in the **S-II** scenario can be seen. Notable is that the optimal values of λ are now negatively correlated with increasing γ. The linear interpolation does not describe the relationship in a proper way. Figure 3.7(b) shows $\sqrt{\text{MISE}_\lambda/\text{MISE}_\infty}$ as a function of $\gamma^{1/3}$, which behaves linearly, just as in the $h-$model.



(a) Optimal values of λ as a function of $\log_{10}\gamma$

(b) $\sqrt{\text{MISE}_\lambda/\text{MISE}_\infty}$ as a function of $\gamma^{1/6}$ together with a linear interpolant

Figure 3.6: Results for the $\lambda-$model in the **S-I** scenario



(a) Optimal values of λ as a function of $\log_{10}\gamma$

(b) $\sqrt{\text{MISE}_\lambda/\text{MISE}_\infty}$ as a function of $\gamma^{1/3}$ together with a linear interpolant

Figure 3.7: Results for the $\lambda-$model in the **S-II** scenario for $k = 3$

Comparing the two models

In both scenarios above for the paint thickness the different performance between the $h-$model and the $\lambda-$model were not that significant. The square root of the ratio of the calculated optimal MISE for the both models for the different cloud factors can be seen in Figure 3.8 in the form $\sqrt{\text{MISE}_h/\text{MISE}_\lambda}$ for the **S-I** and **S-II** scenario with $k = 3$, respectively. The $\lambda-$model did give about 10% less error (in terms of the square root of the MISE), but this method is computationally heavier and more involved to implement in combination with subsequent variations of the kernel estimation model and therefore this minor benefit from the smaller error does not motivate why to use this model instead of the $h-$model for the rest of this thesis, which is easier to combine with following methods. Hence, the rest of the thesis will focus on using the $h-$model along with its modifications.

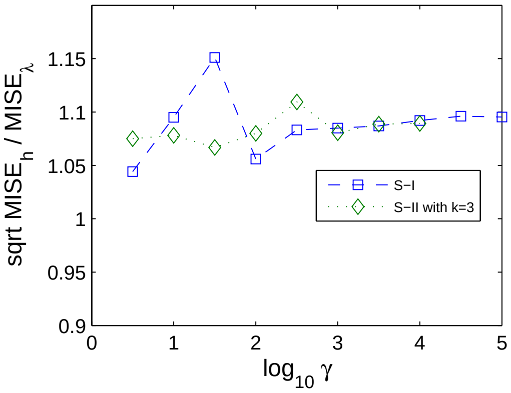
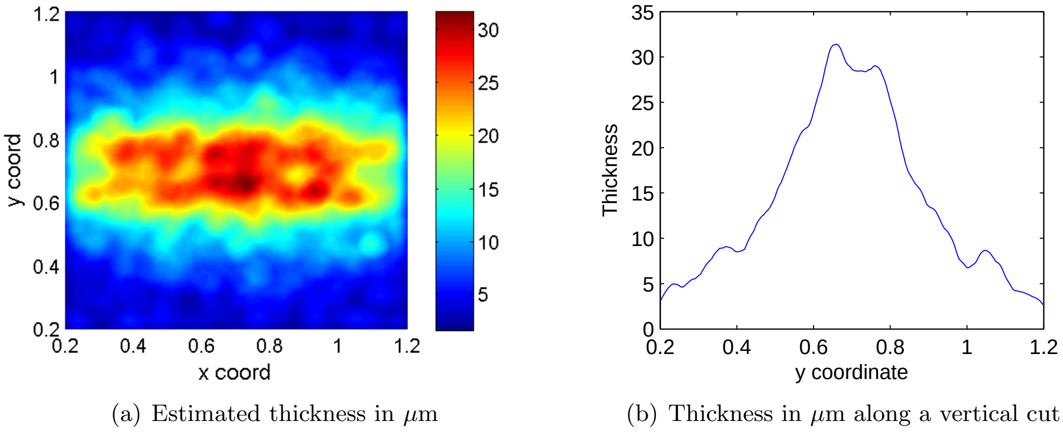


Figure 3.8: The fraction of the optimal MISE for the two different paint thickness distributions as a function of $\log_{10}\gamma$

3.3 Thickness estimations for data generated by IPS

IPS was used to simulate the painting process on a one times one square to resemble the above cases. The resulting estimated thickness from one stroke can be seen in Figure 3.9(a) and the thickness along a vertical line at the center of the plate can be seen in Figure 3.9(b). In this simulation 80 000 droplets per second during the total time of 6 seconds were simulated, which corresponds to a cloud factor of about 16 000 in this specific setup (in IPS the user does not specify the cloud factor directly, but through the number of droplets per simulated second).

As mentioned earlier, a low value of the bandwidth will reduce the bias and increase the variance of the estimated thickness. By using a smaller value of h (14 mm was used) and repeating the painting process several times using a smaller cloud factor and averaging the estimated thickness, a slightly biased reference thickness with low variance will be obtained which will be interpreted as the true thickness. Figure 3.10 shows the



(a) Estimated thickness in μm

(b) Thickness in μm along a vertical cut

Figure 3.9: Estimated paint thickness in μm on the plate with data generated by IPS, using a value of 50 mm for h

obtained reference thickness. Using this thickness, the error in the thickness in each point of a new simulation with a realistic cloud factor can be estimated, and from this the corresponding MISE be obtained by using Equation 3.3. As in the previous scenarios, the MISE was calculated by varying h and using cubic splines to interpolate between the different values. The ISE from 11 different painting simulations were averaged to yield an estimation of the MISE. Figure 3.11 shows the square root of the MISE as a function of h. The optimal value for h was found to be 65 mm. Since the reference thickness is slightly biased, the calculated h of 65 mm is to interpreted as an upper value of the optimal h. However, in means of MISE it is better to use a too large than too small value of h since it grows more rapidly to the left than to the right of its optimum, according to Figure 3.11.

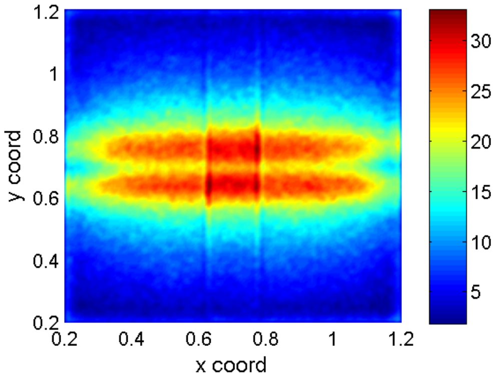


Figure 3.10: Reference thickness in μm, using $h = 14$mm

In Figure 3.12 the resulting estimated paint thickness from five strokes of the painting

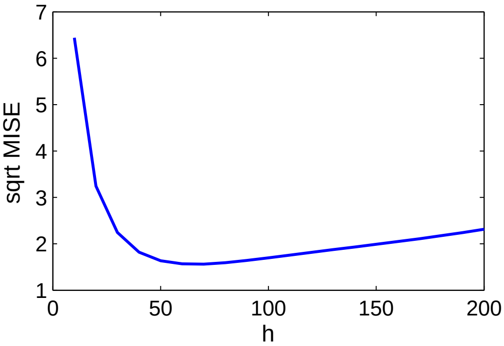


Figure 3.11: Square root of MISE in μm as a function of h in mm for the plate in the **IPS-I** scenario

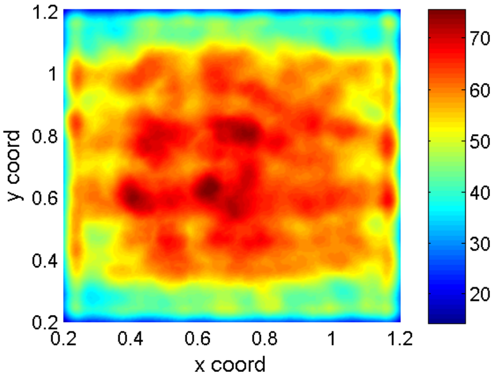


Figure 3.12: Estimated thickness μm for the **S-II** scenario

robot, the **IPS-II** scenario, can be seen using $h = 55$ mm.

Paint thickness estimations for a Volvo V60

Finally a three dimensional object, a part of a Volvo V60, was painted using IPS. Figure 3.13 shows the painting process in action. The major difference in the estimation of paint thickness compared to the two dimensional surfaces used before is that it is not anymore trivial how to calculate the distance on the surface. In [8] this was handled by assuming a locally flat surface when performing the thickness estimate and project nearby droplets on this plane through their impact velocities. On this plane the distances are then easily calculated as the Euclidean distances. However, for simplification, in this thesis the three dimensional Euclidean distance were used to find the distance from an impact to a point $\boldsymbol{x} \in S$, not including the effects of a possibly curved surface. Figure 3.14 shows a thickness estimation on the Volvo V60 using the data obtained from the IPS simulation.

It is clear that the estimates work good for domains of the surface where the curvature

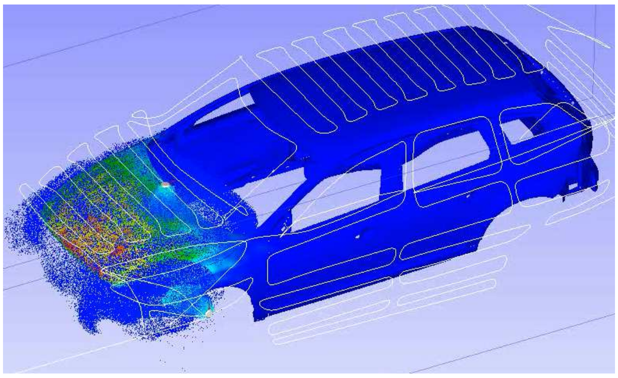


Figure 3.13: Simulation the painting process of a VolvoV60 in IPS (geometry courtesy of Volvo Car Corporation)

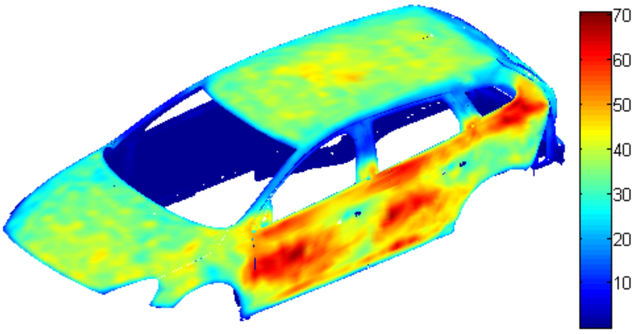


Figure 3.14: Estimated thickness of a Volvo V60, measured in μm (geometry courtesy of Volvo Car Corporation)

is very low and that are not very close to the edges. However, along the side the estimates performs worse. Because of this, the thickness were estimated considering only a subset of the object, yielding the result presented in Figure 3.15. The average estimated thickness in this figure was about 32 μm.

3.4 Anisotropical methods for estimating the thickness

From Figure 3.3 it is clear that there is a tradeoff between the variance and bias of an estimate. This is because the support of the kernel needs to be hold fairly wide to reduce noise and hence leads to biased estimates. One possible way to reduce both the bias and the variance of the estimate simultaneously is to adapt the kernel so that it no longer necessarily has a circular support. In the **S-II** scenario, the only actual variation (*i.e.* that is not due to chance) is in the $x-$direction. Hence, the wider the support of the kernel along the $x-$axis, the larger the bias. On the other hand, in the $y-$direction the thickness is on average the same, hence incorporating more points in the $y-$direction

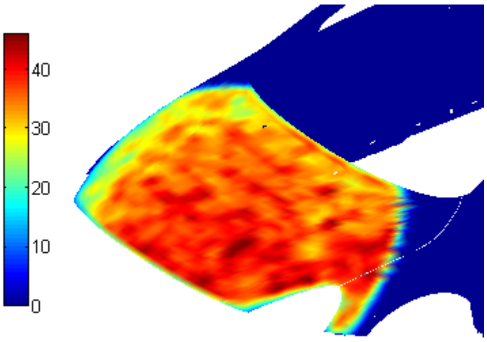


Figure 3.15: Estimated thickness of a part of the Volvo V60, measured in μm (geometry courtesy of Volvo Car Corporation)

will decrease the variance of the estimation, while not affecting the bias. In Figure 3.16 two estimates of the thickness along a horizontal cut in the $x-$direction are shown, using a cloud factor of 100, the value 58 mm for the bandwidth h and a value of $k = 3$. The kernel is adjusted to incorporate impacts twice as distant in the $y-$direction but only half in the $x-$direction compared to the native version (using just a circular neighborhood). Note the reduction in bias around the peaks of the paint thickness using the anisotropic method.

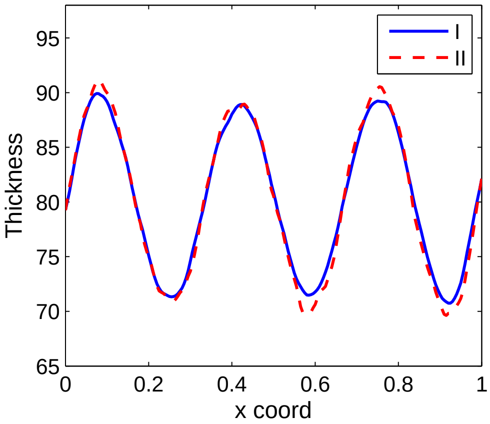
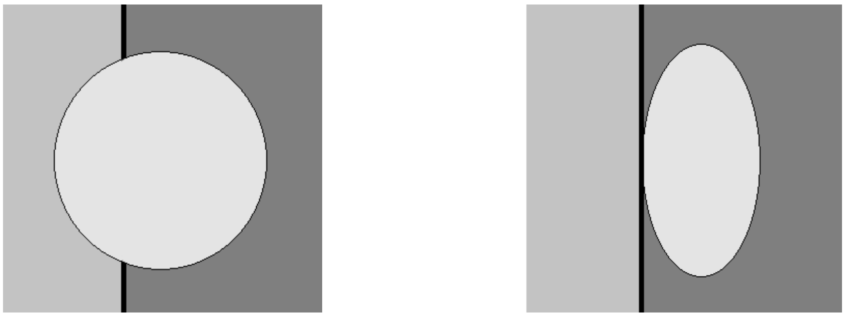


Figure 3.16: Original kernel estimate (**I**) and kernel estimate adjusting the kernel (**II**), measured in μm

As can be seen in Figure 3.16, using anisotropical methods for the kernel estimates (which means to make the support of the kernel directional dependent), can improve the performance of the estimations. It is however not always possible to on before hand choose how to adjust the kernel. Methods for automatically chose how to adjust the kernel is therefore needed. Next follows a description of an anisotropical method based on [16, 18].

To begin, the kernel estimation method with isotropic support (that is, the same in all directions) but with a larger value of the bandwidth (to reduce the impact of noise) is used to try to locate points near the boundaries or with irregularities in the paint thickness (that is not due to the randomness in the process itself). Calculating the gradient of the estimated thickness in each of the corresponding grid points, $\nabla\widehat{T}(\boldsymbol{x};\gamma,h)$, these points will result in larger values of the gradient than in the inner points with a more evenly spread thickness. This both gives the information that the thickness vary significantly and also in which direction this variation occurs. From this the kernel can be modified so that the support in the direction of most variation will decrease, see for instance Figure 3.17 for an example how this will, in theory, optimally adopt the support of the kernel at the boundaries.



(a) Isotropic (b) Anisotropic

Figure 3.17: An example of isotropic and anisotropic kernel support near an edge

More technically [16, 18], from the calculated gradient $\nabla\widehat{T}(\boldsymbol{x};\gamma,h)$, a *structure tensor* **S** is created by forming the product

$$\mathbf{S} = \nabla\widehat{T}\nabla\widehat{T}^T.$$

S is a positive semidefinite symmetric matrix, hence [12] it is orthogonally diagonalizable, *i.e.* there exist matrices **Q** of eigenvectors of **S** and

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix},$$

of real eigenvalues such that $\mathbf{D} = \mathbf{Q\Lambda Q}^T$. From **S** a *diffusion tensor* **D** is then created by forming the matrix product

$$\mathbf{D} = \mathbf{QMQ}^T,$$

with

$$\mathbf{M} = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix}$$

where $\mu_1 = 1$ and

$$\mu_2(\lambda) = \frac{1}{1+\left(\lambda/q\right)^{\alpha}}, \tag{3.4}$$

with λ being the (nonnegative) eigenvalue of $\mathbf{S}$. q and α are parameters used for calibrating the model. The parameter q is used to determine how much variation is needed for the model to not consider the variation as noise (that should not be incorporated in the model) and α controls the steepness, see Figure 3.18 for an illustration of the function $\mu_2(\lambda)$.

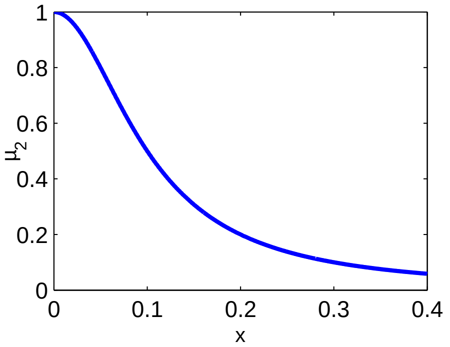


Figure 3.18: $\mu_2(x)$ for $q = 0.1$ and $\alpha = 2$

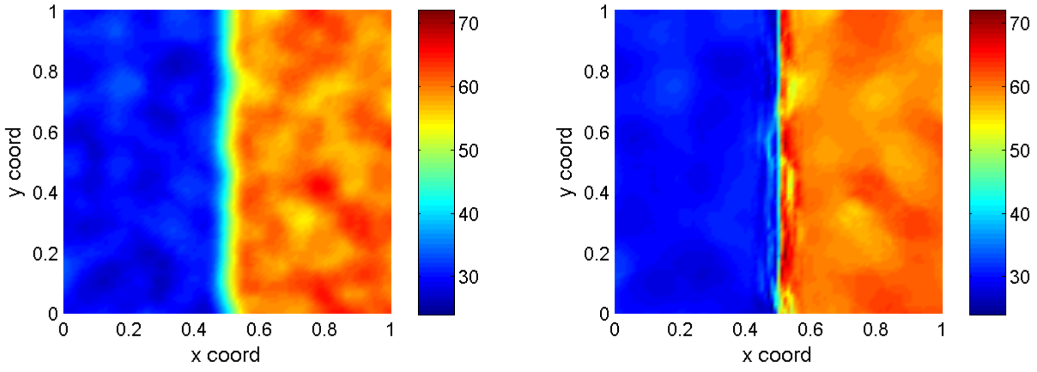
Using the matrix $\mathbf{D}$, the weighted distance from the point $\boldsymbol{x}$ to droplet number i is calculated as $\sqrt{(\boldsymbol{x}-\boldsymbol{X_i})^T\mathbf{D}^{-1}(\boldsymbol{x}-\boldsymbol{X_i})}$. This yields the anisotropic kernel estimation

$$\begin{aligned}\widehat{T}_D(\boldsymbol{x};\gamma,h) =& \frac{\gamma}{h^2\sqrt{\det \mathbf{D}}}\sum_{i=1}^{n} V_i K\left(\frac{\sqrt{(\boldsymbol{x}-\boldsymbol{X}_i)^T\mathbf{D}^{-1}(\boldsymbol{x}-\boldsymbol{X}_i)}}{h}\right)\\ =& \frac{\gamma}{h^2\sqrt{\det \mathbf{D}}}\sum_{i=1}^{n} V_i K\left(\frac{\sqrt{(\boldsymbol{x}-\boldsymbol{X}_i)^T\mathbf{QM^{-1}Q^T}(\boldsymbol{x}-\boldsymbol{X}_i)}}{h}\right).\end{aligned} \tag{3.5}$$

By using Equation 3.5 to adapt the kernel anisotropically, it should be possible to reduce the bias near the step in the paint thickness for the **S-I** scenario. In Figure 3.19 the estimated thickness for both the isotropic and anisotropic method based on one realization with $\gamma = 10^3$, can be seen when only considering inner points.

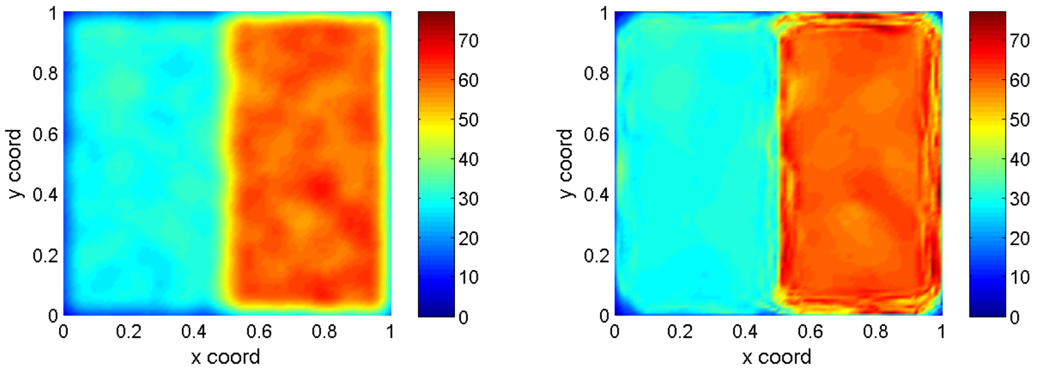
For the isotropic model the value of h calculated above that minimized the MISE was used and in the anisotropic model a 50% larger h was used. Also, in Figure 3.20 a similar estimation was made but now also incorporating the boundary effects and Figure 3.21 shows a horizontal cut of the corresponding estimate. The parameters q and α were chosen to $q = 10^{-8.5}$ and $\alpha = 1.5$ respectively. By optimizing with respect to them, the anisotropic method could probably perform a little bit better.

From Figures 3.19, 3.20 and 3.21 it is clear that the anisotropic method seems to handle the step in a much better way than the isotropic method, with a significant reduction in the bias. This is also the results near the boundaries. Also, since a larger



(a) Isotropic (b) Anisotropic

Figure 3.19: Estimated thickness in μm using isotropic respective anisotropic methods, excluding the effects of the boundaries in the **S-I** scenario



(a) Isotropic (b) Anisotropic

Figure 3.20: Estimated thickness in μm using isotropic respective anisotropic kernel, including the effects of the boundaries in the **S-I** scenario

value of h could be used than in the isotropic method, the resulting estimated thickness for points not close to the step or the boundaries were smoother. However, apart from being more computationally demanding than the isotropic method, the anisotropic method also leads to higher variance of the estimated thickness at points where the method adjusts the support of the kernel. This is because the elliptic support of the kernel at these points has a smaller area, and hence will incorporate a fewer number of droplets when performing the estimates. Again there is a trade-off between bias and variance, but over all the anisotropic method seems to work well for this case.

However, when using this method to estimate the paint thickness on the plate with data from IPS, the benefits are not that obvious. This is probably because the intensity of the droplets along the surface has a smoother spatial variation, hence when estimating

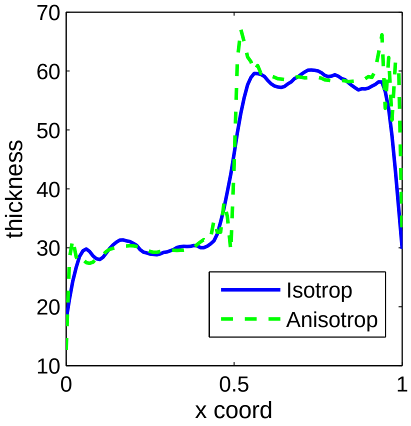


Figure 3.21: Thickness in μm for a horizontal cut for the isotropic and anisotropic thickness estimates in the **S-I** scenario

the thickness at a point $\boldsymbol{x} \in S$, the variations of the intensity locally are almost linear. Hence, the isotropic method will not produce much bias due to this since the variations cancels out. Therefore, using the anisotropic method for this case does not yield much advantage compared to the isotropic. In fact it could perform worse in points with a large gradient since it increases the variance. Figure 3.22 shows a horizontal cut for the **IPS-I** scenario using the isotropic and anisotropic method, with h chosen as 80 and 60 mm respectively.

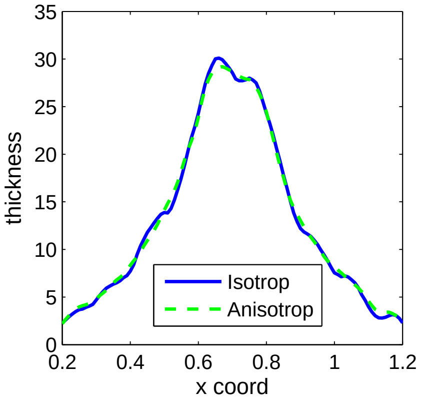


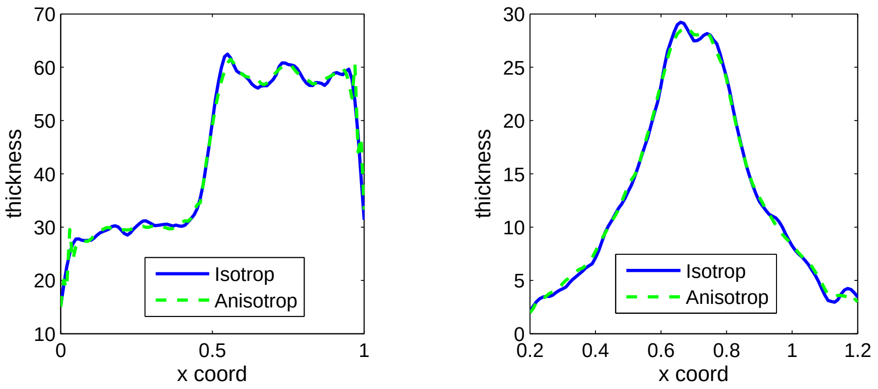
Figure 3.22: Thickness in μm for a horizontal cut for the isotropic and anisotropic thickness estimates in the **IPS-I** scenario

What is problematic and needs to be handled is if the intensity has for instance a local maximum or minimum, as in Figure 3.9(b). When estimating the thickness at or at least very close to the optimum, the estimates will be biased; they will systematically underestimate the thickness at a maximum while overestimating it at a minimum. More generally, the estimates will be underestimate at points where the thickness is concave and overestimated where it is convex. In these points it should be better to use higher order derivatives than only the gradient of the intensity when adopting the support of the kernel, here we will limit ourselves to second order derivatives. If the second order derivatives along one direction is of much greater magnitude than in its orthogonal direction, this should be incorporated when adjusting the support. It is also reasonable to put a limit on the ratio between the two semi-axes of the elliptic support, since the gain of some reduction in the bias is not worth the cost of a greatly increased variance. This is easily done by choosing a lower limit for μ_2 in Equation 3.4; the ratio between the two semi-axes of the elliptic kernel support is $\sqrt{\mu_1/\mu_2}$.

To let second order derivatives of the estimated thickness also be incorporated when adjusting the kernel, the Hessian matrix **H**, consisting of approximate second order derivatives of the estimated thickness using the isotropical method, will be used by choosing

$$\mathbf{S} = \mathbf{H}$$

in Equation 3.5 and using the absolute value of the eigenvalues of **H** in Equation 3.4 when calculating the matrix **M**; this is important since **S** will no longer necessary be a positive semidefinite matrix. It will however be symmetric, and hence have real eigenvalues [12]. Using the anisotropic method based on the Hessian and with a ratio limit between the semi-axes for the elliptical support of the kernel produces the results in Figures 3.23(a) and 3.23(b) for the **S-I** and **IPS-I** scenario respectively, showing the thickness along a horizontal cut of the plates.



(a) **S-I** scenario

(b) **IPS-I** scenario

Figure 3.23: Thickness in μm for a horizontal cut for the isotropic and Hessian based anisotropic thickness estimates using a ratio limit

Neither of the anisotropic methods did improve the thickness estimates much for the **IPS-I** scenario. For the **S-I** scenario the gradient based method gave better results than the Hessian based, which gave similar results as the isotropic. The anisotropic method based on the Hessian is however much more sensitive to the parameters used, but optimizing it with respect to them is out of the scope of this thesis.

It could also be possible to combine the two anisotropic methods by introducing a parameter $\alpha \in [0,1]$ and letting

$$\tilde{\mathbf{S}} = (1-\alpha)\ \nabla\widehat{T}\nabla\widehat{T}^T + \alpha\sqrt{\mathbf{H}^T\mathbf{H}},$$

where the square root of the matrix $\mathbf{H}^T\mathbf{H}$ is defined according to [19]. The matrix $\tilde{\mathbf{S}}$ obtained through this procedure resembles the matrix $\mathbf{S}$ used above as it shares its symmetrical properties. Hence, by similar reasoning as above we can construct a matrix $\tilde{\mathbf{D}}$ to be used for calculating (weighted) distances. This yields a anisotropic kernel estimation based on both gradients and second order derivatives. Generalizations to other properties to trigger the reshaping of the kernel are also possible. The above mentioned method and other generalizations of it will however not be used in this thesis.

3.5 Compensation for boundaries

At the corners of the plate the anisotropic method can not successfully be used since it is not possible to adjust the support elliptically so that it mostly falls on the plate. Also, the anisotropic methods did not handle the boundaries in a satisfying way. Other methods are therefore needed to compensate for the part of the kernel not falling on the surface and next two similar methods will be tested for reducing this erroneousness.

The first method to compensate for the lost volume is that for every point $\boldsymbol{x} \in S$ where the thickness is estimated, find out the fraction of the volume of the contribution (which is an elliptic paraboloid centered at $\boldsymbol{x}$ through the diffusion tensor $\mathbf{D}$) of a droplet at $\boldsymbol{x}$ that falls on the surface. Let $g(\boldsymbol{x}), \boldsymbol{x} \in S$ denote this fraction. Then the corrected estimated thickness is obtained via

$$\frac{\widehat{T}(\boldsymbol{x};\gamma,h)}{g(\boldsymbol{x})}. \tag{3.6}$$

Using this method combined with the gradient based anisotropic method yields the following thickness estimation in Figure 3.24 for the **S-I** data, showing again one horizontal cut over the plate, together with the original thickness estimate. Note the smoother behavior of the estimate close to the edges; the thickness is no longer systematically underestimated. Also, Figure 3.25 shows the estimates for the entire surface. It is clear from this figure that the edge compensation algorithm did indeed increase the performance of the estimates along the edges.

Another method, and indeed very similar, to compensate for the edges is that for every droplet's impact location $\boldsymbol{X}_i$ interpolate the gradients or Hessians to obtain a diffusion tensor $\mathbf{D}_i$ for each impact. For every impact location, $\mathbf{D}_i$ can then be used to calculate the paint contribution of droplet i for all nearby points $\boldsymbol{x} \in S$. Also, if the droplet's impact location is close to the boundaries of S, the fraction of the volume falling on the plate is estimated. This fraction is then used to correct the volumes of the

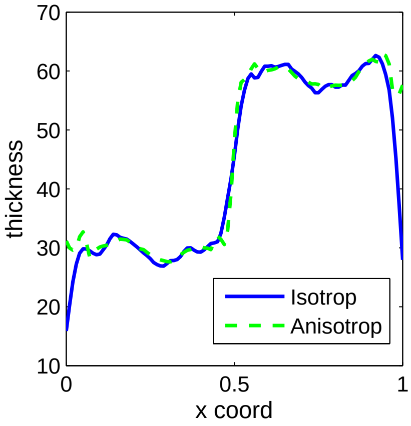
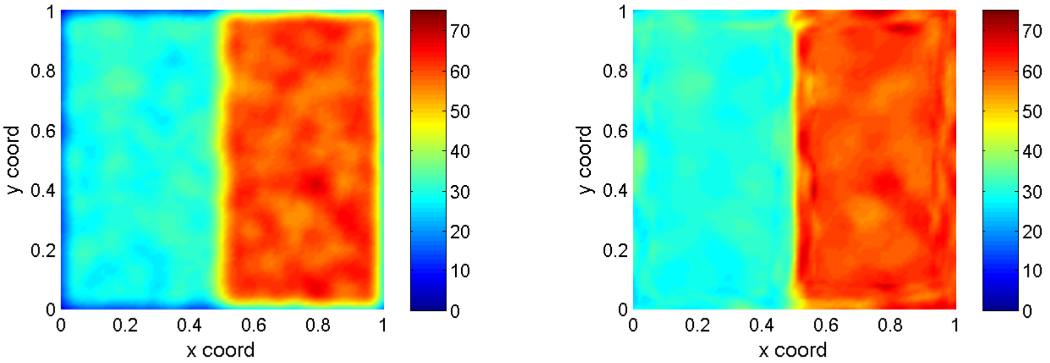


Figure 3.24: Thickness in μm for a horizontal cut for the isotropic and anisotropic thickness estimates in the **S-I** scenario using edge compensation



(a) Isotropic

(b) Anisotropic with edge compensation

Figure 3.25: Thickness in μm for the entire plate in the **S-I** scenario using isotropic and anisotropic with edge compensation

droplets. The advantage with this method over the previous mentioned method is that it (in theory) preserves the total amount of paint applied to the surface. This is a highly desirable property. When using this algorithm on the data in the **S-I** scenario above, the increase in total over all estimated volume was almost 6%. Figure 3.26 shows a horizontal cut for the estimated thickness using this edge compensation algorithm combined with a gradient based anisotropic estimate. It seems as this algorithm tends to overestimate the thickness slightly close to the boundaries.

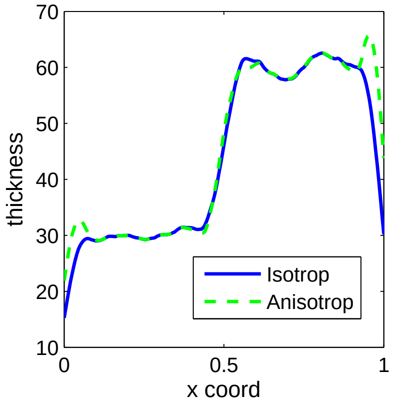


Figure 3.26: Thickness in μm for a horizontal cut for the isotropic and anisotropic thickness estimates in the **S-I** scenario using the second edge compensation method

Calculating the compensation weigths

To compensate for the lost paint when performing estimations close to the edges, for each grid point $\boldsymbol{x} \in S$ where the estimation is performed, we check whether $\boldsymbol{x}$ lies within a distance h from the boundary of the surface. In the case with a two dimensional surface where we know the geometry perfectly, this is easy. For a more general three dimensional surface, this is more involved and needs finer algorithms to work.

For each of the points close to the boundary, a rectangular grid in the $r\phi$ polar coordinate plane containing points $\boldsymbol{z}_i$ are created, corresponding to a circle with radius h in the xy plane . These points are then rotated and stretched by the linear transformation $\sqrt{\mathbf{D}}\ \boldsymbol{z}_i$, which results in a set $\tilde{\boldsymbol{z}}_i$ of elliptically spread points around $\boldsymbol{x}$ [12]. For each point $\tilde{\boldsymbol{z}}_i \in S$, the area of the corresponding elliptical segment is calculated as [12]

$$A = r_i \Delta r \Delta \phi h^2 \sqrt{\det \mathbf{D}},$$

which is then multiplied by the height of the paraboloid at $\tilde{\boldsymbol{z}}_{\boldsymbol{i}} \in S$ obtained through

$$\frac{1}{h^2\sqrt{\det \mathbf{D}}} K_s \left(\frac{\sqrt{(\boldsymbol{x} - \tilde{\boldsymbol{z}}_i)^T \mathbf{D}^{-1} (\boldsymbol{x} - \tilde{\boldsymbol{z}}_i)}}{h} \right).$$

Summing up over all $\tilde{\boldsymbol{z}}_i \in S$ we get an approximate value of $g(\boldsymbol{x})$ which can be used for either of the two edge compensation methods described above through Equation 3.6.

4

Variance estimation

When simulating a painting process, the estimated paint thickness in each point $\boldsymbol{x}$ is a random variable as a consequence of the randomness in the painting simulation. This is both due to the cloud factor $\gamma > 1$ and the randomness when simulating impact locations and the various size of each droplet. A lower γ will result in more simulated points and therefore the variance of the estimated thickness, $\mathrm{Var}(\widehat{T}(\boldsymbol{x};\gamma,.))$, will reduce. There is also another interpretation of variance in this situation, namely that in every simulation, the estimated thickness along the surface will vary and corresponds to $\widehat{T}(\boldsymbol{x};\gamma,.) \in \mathcal{R}^d$ being a multidimensional random variable. This variation will not be our main concern, and for clarification, it will henceforth be called spatial variation. A reasonable assumption would be that the variance and the estimated variance is a function of both $\boldsymbol{x}$ and γ.

Using Equation A.1 from Appendix A for the sample variance directly, is however not applicable in our situation, since we want to be able to estimate the variance after only one painting simulation which, although produces a large amount of impacts, only yields one realization of the estimated paint thickness $\widehat{T}(\boldsymbol{x};\gamma,.)$ at $\boldsymbol{x}$. One could use nearby estimates of the paint thickness and consider these as other realizations of $\widehat{T}(\boldsymbol{x};\gamma,.)$. Very close to $\boldsymbol{x}$ these could often be approximated to be identically distributed as $\widehat{T}(\boldsymbol{x};\gamma,.)$. However, they will be extremely correlated with $\widehat{T}(\boldsymbol{x};\gamma,.)$, hence the assumption of independent realizations is strongly violated and therefore this is not a feasible method for performing variance estimations. Next follows two methods for estimating the variance of the estimated paint. For the rest of the report the isotropic $h-$model, given in Equation 3.1, will be used.

4.1 Regression models

One way to estimate the variance of the paint thickness is to use local multiple linear regression models, and for each point $\boldsymbol{x} \in S$ of interest estimate $\mathrm{Var}(\widehat{T}(\boldsymbol{x};\gamma,h))$ by using nearby points in such a way that the dependence between the estimated paint thickness in these points will be small, but close enough for the assumption of equal variance to still be reasonable. This was successfully done in [8] where the method is used to estimate the variance of the paint thickness on a car fender, both through usage of linear and quadratic models.

In this thesis a similar regression model will be implemented and used in comparison with other models for variance estimation. The model works by finding grid points on circles centered at $\boldsymbol{x}$ with radius chosen large enough to remove most of the correlation in the thickness between the points, but small enough to not include irrelevant points. Using the estimated thickness in these points together with the estimated thickness at $\boldsymbol{x}$, regression models will yield an estimate of the variance. See Appendix A.1 for a presentation of the theory behind the regression analysis; next follows a brief reminder.

The idea behind using regression models is to assume that the paint thickness varies over the surface according to some (unknown) function $Y(\boldsymbol{x})$. Because of the randomness in the process, we can not expect that the paint thickness will be perfectly explained through this relationship, hence a stochastic noise, e is added to the model, yielding

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1} + e,$$

where the x_i's are connected to the coordinates $\boldsymbol{x}$, the β_i's are unknown parameters and the noise e is assumed to be $\mathcal{N}(0, \sigma^2)$ distributed. It is the randomness of e that we want to estimate; hence by estimating the parameters β_i above we can remove the spatial dependence of the estimated thickness, yielding realizations of the noise e from which an estimate $\widehat{\sigma}$ of σ can be obtained.

Both linear as well as quadratic models will be used. Using the linear model to estimate the variance of the estimated paint thickness, the effect of the nonlinear spatial variations will not be corrected for and fully incorporated in the estimated variance and hence lead to an overestimation. The advantage of using a quadratic model is that the spatial variations of the estimated thickness will be excluded (to some extent) in the estimated variance. The linear model that will be used is

$$Y = \beta_0 + \beta_1 x' + \beta_2 y' + e, \tag{4.1}$$

where x' and y' are coordinates relative to $\boldsymbol{x}$. Similarly, the quadratic model that will be used is of the form

$$Y = \beta_0 + \beta_1 x' + \beta_2 y' + \beta_3 x'^2 + \beta_4 y'^2 + \beta_5 x' y' + e. \tag{4.2}$$

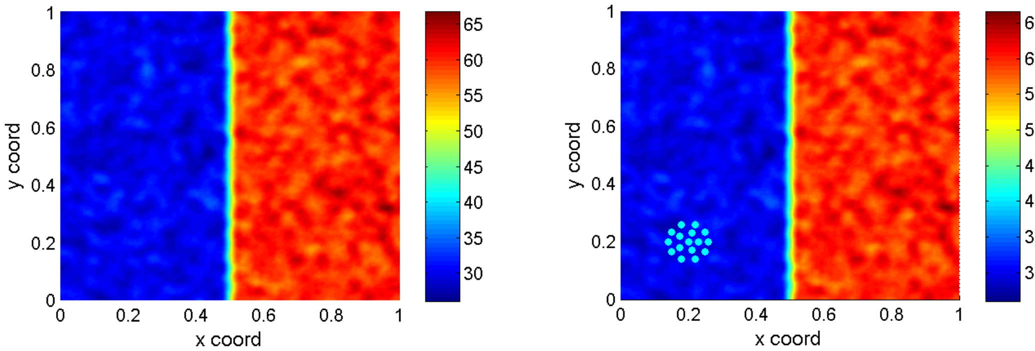
For the regression to perform well, many points should be included when estimating the variance. This could be accomplished by including extra layers of points to be used in the regression. If the thickness is roughly the same over a large part of the surface around the point in consideration, this is reasonable. However, when estimating the variance at parts of the surface with high spatial variations of the thickness, the part of the surface containing the points used in the regression has to be kept small to not contradict the necessary assumptions of the regression analysis; *i.e.* that the thickness can be explained by the functions described above. The low number of points used may decrease the accuracy of the regression, leading to large errors in the estimated variance.

The S-I scenario

The regression models were used to estimate the standard error of the estimated thickness for the **S-I** scenario by using the $h-$model. A cloud factor of 100 and $h = 31$ mm, the h found to minimize the MISE, was used for the thickness estimation. In Figure 4.1(a) the estimated thickness can be seen. Linear and quadratic regression models were used to estimate the variance at the point $[0.2, 0.2]$. Figure 4.1(b) shows this and nearby points, indicated by dots in the lower left part of the plate, that were used in the regression. Since the true variance is unknown it is not possible to use it for comparing with the obtained estimates; however, by simulating the process hundreds of times, a more accurate estimate can be obtained by means of the sample variance. Using the estimates $\widehat{\sigma}$, the *coefficient of variation*, defined as

$$c_v = \frac{\widehat{\sigma}}{\widehat{T}(\boldsymbol{x}; \gamma, h)},$$

can be obtained. In Table 4.1 the results are summarized, based on 500 simulations.



(a) Estimated thickness

(b) Points used in the regression

Figure 4.1: Estimated thickness (μm) used for the the regression together with the regression points indicated by dots

Table 4.1: Estimated $\widehat{\sigma}$ (μm) for the regression models and through simulations for the **S-I** scenario together with the coefficient of variation c_v (%)

	Linear regression	Quadratic regression	Simulations
$\hat{\sigma}$	1.12	1.04	1.12
c_v	3.51	3.29	3.73

The regression using the linear model gave the same estimated standard error as the sample variance obtained through the simulations. There is however one major drawback with the linear model, namely that it will overestimate the standard error if applied to points where the underlying paint intensity has a nonlinear spatial variation. This would be the case if trying to estimate the standard error for the estimated thickness close to the step or in other points where there is a large variation in the thickness. Using regression to estimate the errors close to the boundaries is also problematic. If some regression points falls outside the surface they must of course be discarded and hence the regression will use less points than it should. The linear model requires at least 4 points to produce an estimate of σ (else the model will fit the data exactly; the system of equations produced through Equations 4.1 and 4.2 needs to be overdetermined [14]), while the quadratic requires 7 or more. It is therefore not possible to use this method on the entire surface.

The IPS-I scenario

Regression models were also used to estimate the variance in the estimated thickness at the center of the plate in the **S-I** scenario with the data used in Figure 3.9(a), but instead using a value of $h = 55$ mm for the thickness estimation. Figure 4.2 shows the estimated thickness again together with the points that was used in the regression.

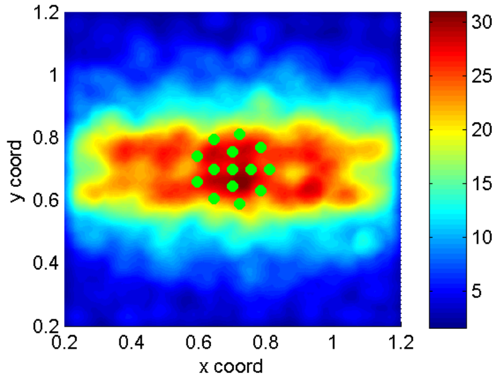


Figure 4.2: Thickness (μm) used in the regression together with the points used for the **IPS-I** scenario

To evaluate the accuracy of the estimates, a total of 105 independent and identical painting simulations in IPS were produced. Using these, an estimate of the standard error for the center of the plate were found by means of the sample standard error. In Table 4.2 the results of the regression and simulations are summarized. As expected, the linear model resulted in a larger estimate than the quadratic, as it does not correct for any nonlinear spatial variation in the estimated thickness before performing the estimation. The simulations produced estimates which lies between those obtained through the two regression models.

Table 4.2: Estimated $\widehat{\sigma}$ (μm) for the regression models and through simulations for the center of the plate in the **IPS-I** scenario together with the coefficient of variation c_v (%)

	Linear regression	Quadratic regression	Simulations
$\hat{\sigma}$	2.43	1.22	1.42
c_v	8.60	4.30	5.52

The models were also used to estimate the standard error in the estimated thickness for the whole plate simultaneously. In Figure 4.3 the estimates $\widehat{\sigma}$ and the coefficient of variation using the linear and quadratic model respectively, can be seen. Since the boundaries are problematic, especially for the quadratic model, points directly at the boundaries are not considered when performing the regression (the reason for the strange behavior of the estimates along the boundaries of the figures is because some of the regression points falls outside of the surface). In the figures there is an upper bound on the colorbars, set to 3 μm for the error and 30% for the coefficient of variation and the reason for the upper limits is because the estimates are noisy; hence some outliers are to be expected. These are however not of our primary interest, thus white means that the estimations lie at or above these bounds. Using the 105 painting simulations, estimates of the standard error for the whole plate were obtained and in Figure 4.4 $\widehat{\sigma}$ and c_v can be seen.

To evaluate the obtained results, the relative error η (considering the results obtained from the simulations to be the correct values, denoted by σ below) were calculated, excluding the points in Figure 4.3 close to the boundaries. The relative error is defined as [6]

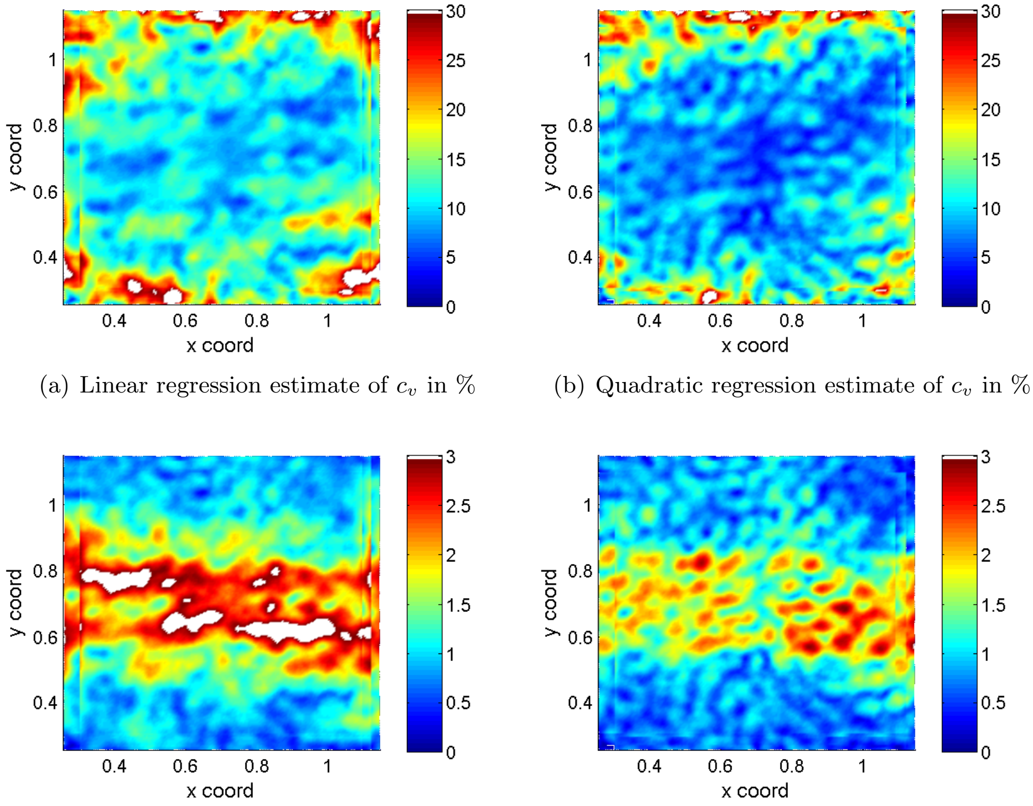
$$\eta = \frac{|\widehat{\sigma} - \sigma|}{\sigma}.$$

Subsequently the relative error taken with sign,

$$\eta_s = \frac{\widehat{\sigma} - \sigma}{\sigma},$$

will also be used and denoted *signed* relative error. The reason for including the sign is to detect if the estimates are systematically over or underestimated. A positive value of η_s means that the model overestimates σ.

A kernel density estimate for the distribution of the signed relative error, η_s, obtained from the linear model for all points can be seen in Figure 4.5(a). Figure 4.5(b) shows the similar kernel density estimate for the quadratic model. The mean, median and upper 95% values of the (unsigned) relative error obtained are summarized in Table 4.3.



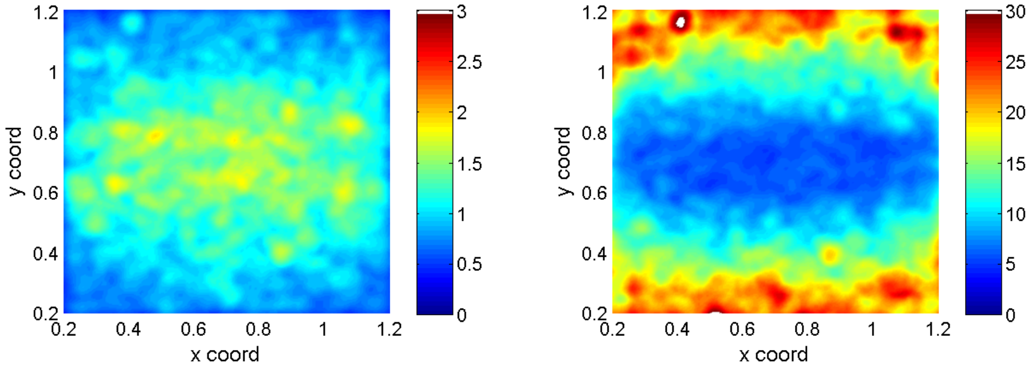
(a) Linear regression estimate of c_v in %

(b) Quadratic regression estimate of c_v in %

(c) Linear regression estimate of $\widehat{\sigma}$ in μm

(d) Quadratic regression estimate of $\widehat{\sigma}$ in μm

Figure 4.3: Linear and quadratic regression for the **IPS-I** scenario



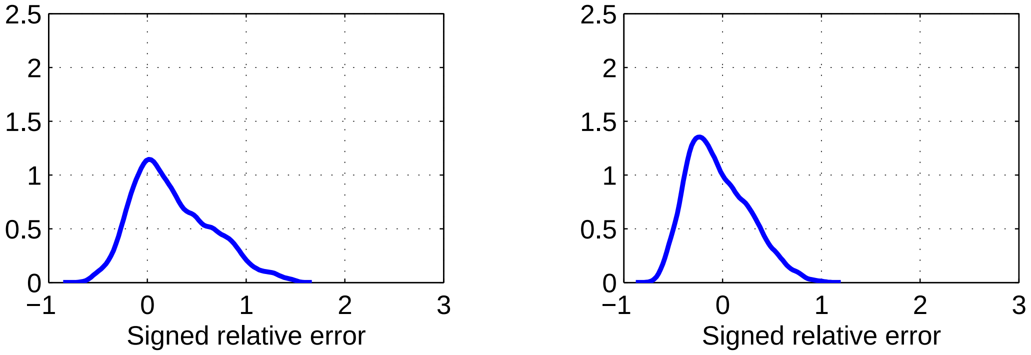
(a) Sample estimate $\widehat{\sigma}$ in μm, based on 105 painting simulations

(b) Sample c_v in %, based on 105 painting simulations

Figure 4.4: The sample standard error and coefficient of variation obtained from 105 painting realizations in the **IPS-I** scenario

Table 4.3: Relative error for the **IPS-I** scenario using regression

	Linear regression	Quadratic regression
Mean	0.35	0.26
Median	0.26	0.24
Upper 95%	0.98	0.58



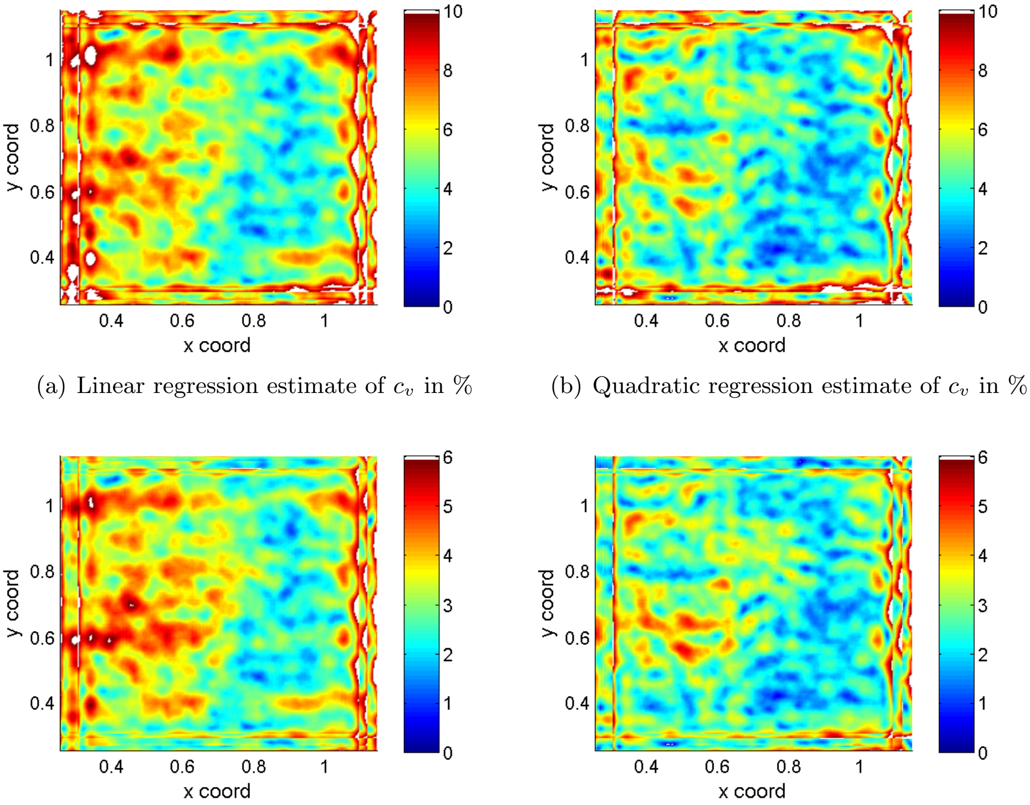
(a) Linear model

(b) Quadratic model

Figure 4.5: Kernel density estimate for the distribution of the (signed) relative error η_s of $\widehat{\sigma}$ for the linear and quadratic models using regression

The IPS-II scenario

Figure 4.6 shows results obtained via regression for the plate painted with five strokes. The mean standard error obtained were 2.69 μm and 2.24 μm for the linear and quadratic model respectively.



(a) Linear regression estimate of c_v in %

(b) Quadratic regression estimate of c_v in %

(c) Linear regression estimate of $\widehat{\sigma}$ in μm

(d) Quadratic regression estimate of $\widehat{\sigma}$ in μm

Figure 4.6: Linear and quadratic regression for the **IPS-II** scenario

4.2 Estimation using bootstrap

Another way to estimate $\mathrm{Var}(\widehat{T}(\boldsymbol{x};\gamma,h))$ from only one simulation of the painting process is to use bootstrap methods to mimic the random behavior of the painting process. This is a more computationally demanding method than the regression and works as follows. Given a set A of impacts (either generated by IPS or synthetic data), we create a new set B of impacts by sampling with replacement from A. From B the paint thickness is estimated. Repeating this N (often thousands) times yields N estimates of $\widehat{T}(\boldsymbol{x};\gamma,h)$ for each $\boldsymbol{x} \in S$ and hence $\mathrm{Var}(\widehat{T}(\boldsymbol{x};\gamma,h))$ can be estimated by means of the sample variance. For more theory regarding bootstrap, see Appendix A.1

It could also be reasonable to let the size of B, the re-sampled impacts, be a stochastic variable as the number of droplets that will hit the surface differs from each simulation. If every droplet has a probability p of hitting the surface, this would imply that the size of B is $\mathrm{Bin}(n,p)-$ distributed with n being the number of droplets leaving the painting robot. For the data obtained through IPS, p can easily be estimated as [14]

$$p = \frac{n}{\#\text{ droplets hitting the surface}},$$

but for the synthetic data this is not a possible method since the only accessible information is the product np. Fortunately, for n large and p small, the number of droplets is approximately $\mathrm{Po}(np)-$ distributed because of the Poisson approximation of the Binomial distribution [14]. Hence, instead of letting the size of B be Bin(n,p), we can let it be Po(np) distributed [5].

Using this methodology is especially efficient when performing bootstrap at a single point $\boldsymbol{x} \in S$. First all impacts within the support of the kernel are located and from these the fraction ρ of the total of the m impacts falling within the support of the kernel at $\boldsymbol{x}$ can be estimated. By sampling Bin(m,ρ) or Po($m\rho$) number of droplets from this set in each re-sampling procedure, we incorporate more randomness of the painting process, while only having to consider those droplets that will influence our estimate.

The S-I scenario

First we use bootstrap to produce an estimate of the variance of the estimated thickness at the plate at $\boldsymbol{x} = [0.2,\ 0.2]$ in the **S-I** scenario; the same point as for the regression models using the same data set. The above described procedure for estimation at a single point was used and in Figure 4.7 the estimated probability density of the thickness can be seen, obtained through one dimensional Gaussian kernel density estimation. The standard error was calculated to $\widehat{\sigma} = 1.18$ μm and the coefficient of variation to $c_v =$ 3.7%.

The IPS-I scenario

Bootstrap estimates of the standard error for the estimated thickness were also produced for the plate painted with IPS from Figure 4.2 using the same value of h. Figure 4.8(b)

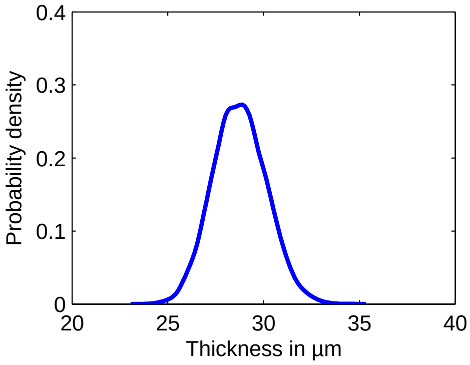
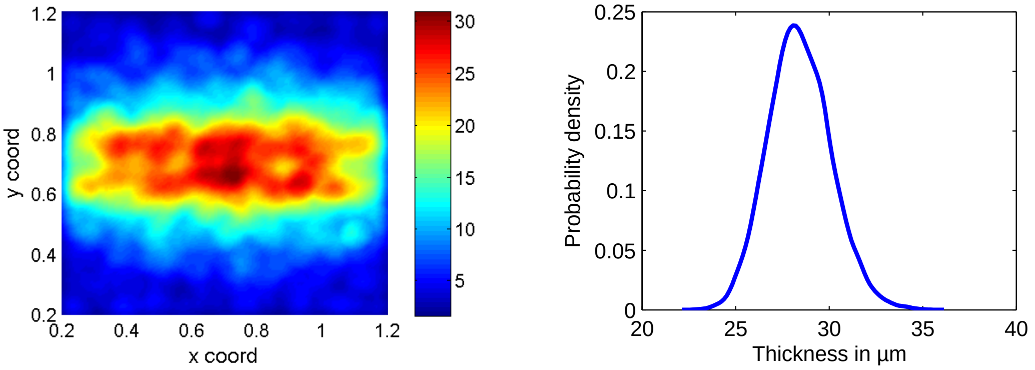


Figure 4.7: Kernel density estimate of $\widehat{T}(\boldsymbol{x};\gamma,h)$ at $\boldsymbol{x} = [0.2, 0.2]$ obtained from 10 000 bootstrap re-samples for the **S-I** scenario

shows the bootstrap distribution of the thickness at the center of the plate in Figure 4.8(a), obtained through Gaussian kernel density estimation. The standard error was computed to 1.67 μm and the coefficient of variation to $c_v = 5.87\%$, which falls between the values obtained by linear and quadratic regression and slightly above the value obtained from the 105 simulations.



(a) Estimated thickness in μm

(b) Bootstrap distribution of the thickness at the center of the plate, obtained from 10 000 bootstrap re-samples

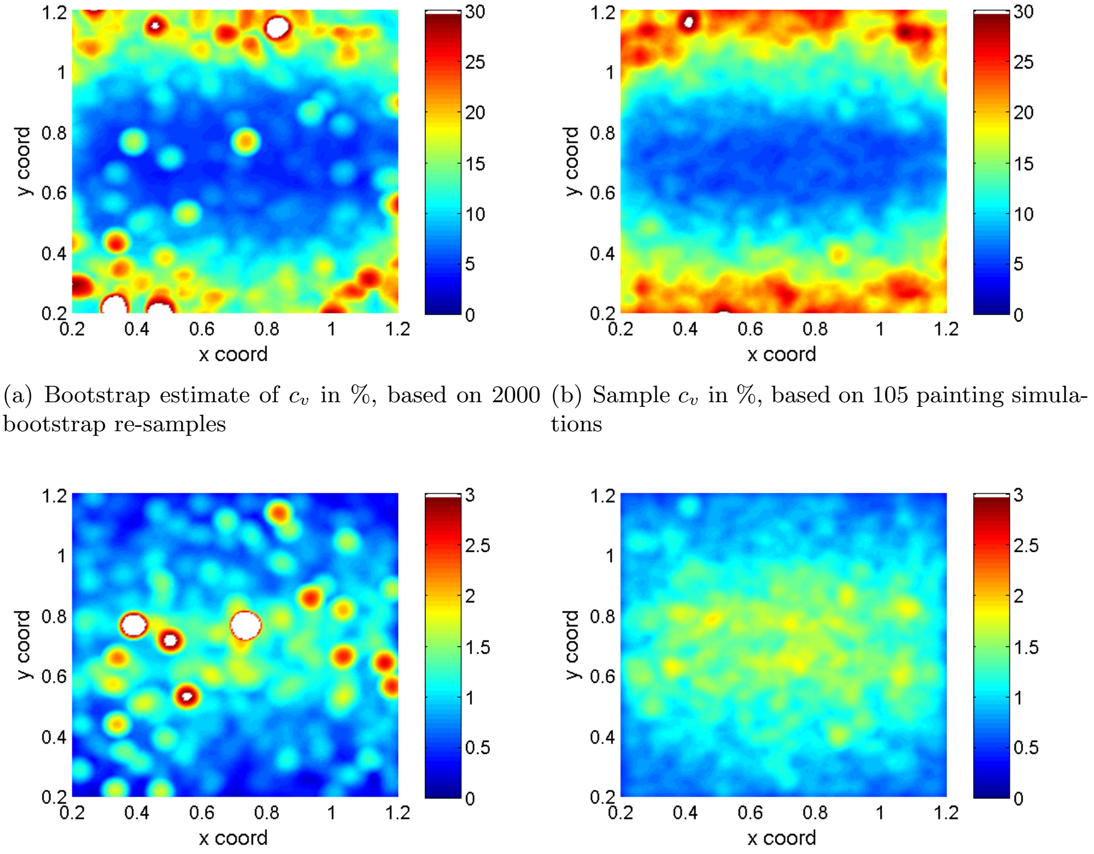
Figure 4.8: Estimated paint thickness on the plate together with the corresponding bootstrap distribution of the thickness at the center of the plate

As in the analysis using the regression models, bootstrap was used to estimate the standard errors all over the plate. Figure 4.9(a) shows the estimated coefficient of variation, obtained through bootstrap, for the whole plate using $h = 55$ mm. Also here there is a cut off at 30% in the figure, and all values above this are colored white. In Figure 4.9(c) the bootstrap estimates of the standard error can be seen with a cut off at 3 μm. These are to be compared to Figure 4.9(b) which shows the c_v based on the sample variance from the 105 independent painting simulations and Figure 4.9(d) showing the estimated $\widehat{\sigma}$, for the same cloud factor and value of h. Figure 4.10 shows a kernel density estimate of the signed relative error η_s obtained from bootstrap (using only the points which were included in the regression models to make the estimates comparable). To verify the consistency of the bootstrap estimates in the above case, the method was also tested on new verification data obtained through two new paint simulations. In Figure 4.12 the distribution of the corresponding signed relative errors for the two new data sets are presented. All results are summarizied and presented in Table 4.4.

Table 4.4: Relative error for the **IPS-I** scenario using bootstrap

	Original data set	First verification data set	Second verification data set
Mean	0.27	0.27	0.22
Median	0.23	0.22	0.19
Upper 95%	0.59	0.65	0.51

From Table 4.4 it can be seen that bootstrap produces consistent estimates for the three data sets investigated; some variations in the estimates are to be expected.



(a) Bootstrap estimate of c_v in %, based on 2000 bootstrap re-samples

(b) Sample c_v in %, based on 105 painting simulations

(c) Bootstrap estimate of $\widehat{\sigma}$ in μm, based on 2000 bootstrap re-samples

(d) Sample estimate $\widehat{\sigma}$ in μm, based on 105 painting simulations

Figure 4.9: Bootstrap estimate of $\widehat{\sigma}$ and c_v for the **IPS-I** scenario together with estimates obtained from 105 painting realizations

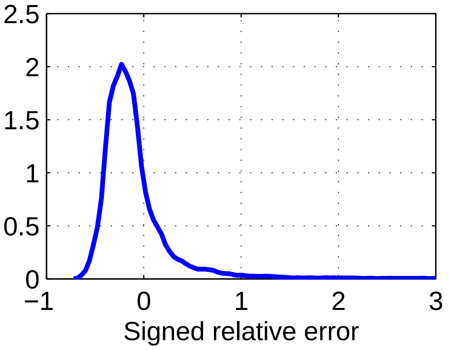
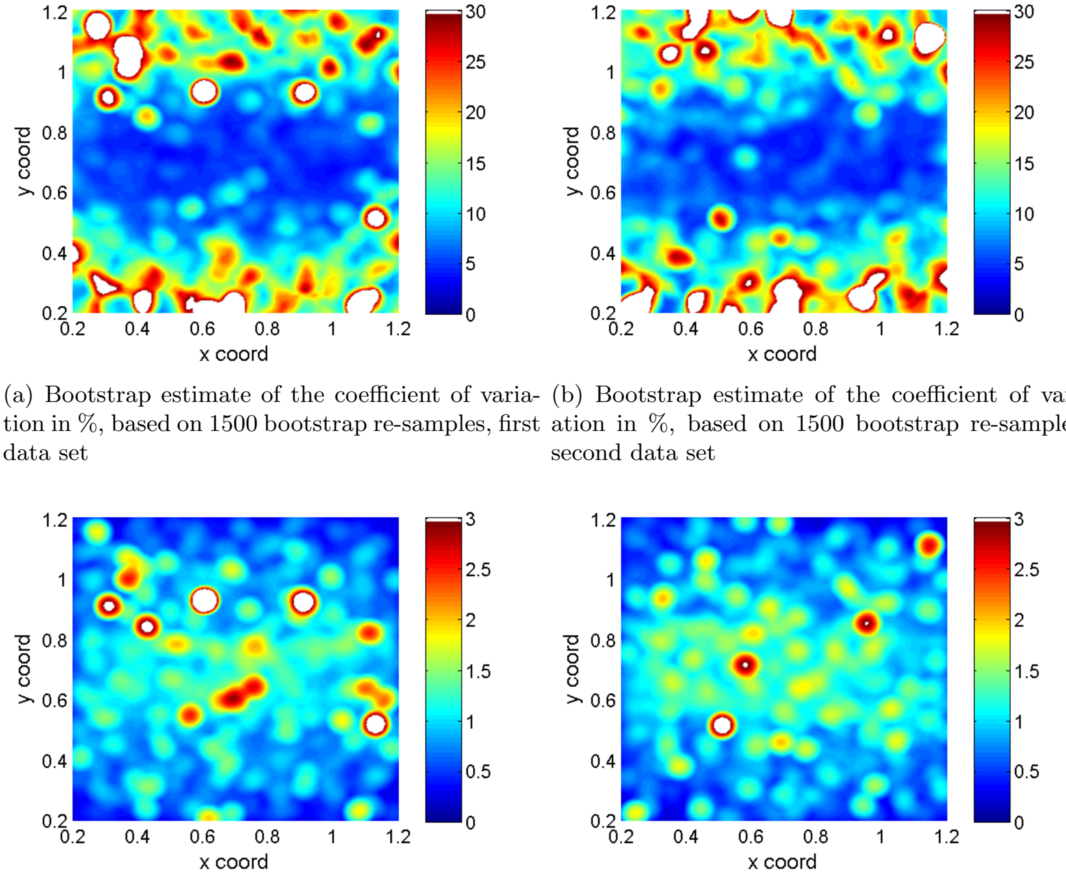


Figure 4.10: Kernel density estimate for the distribution of the (signed) relative error η_s of $\widehat{\sigma}$ using bootstrap in the **IPS-I** scenario



(a) Bootstrap estimate of the coefficient of variation in %, based on 1500 bootstrap re-samples, first data set

(b) Bootstrap estimate of the coefficient of variation in %, based on 1500 bootstrap re-samples, second data set

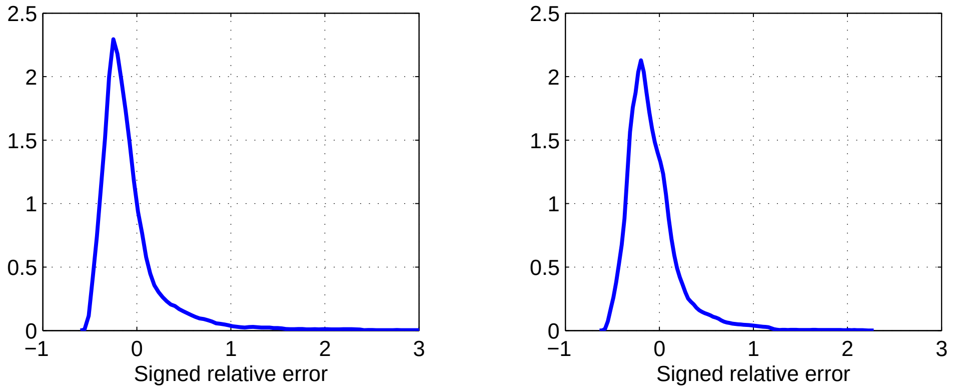
(c) Bootstrap estimate of the standard error in μm, based on 1500 bootstrap re-samples, first data set

(d) Bootstrap estimate of the standard error in μm, based on 1500 bootstrap re-samples, second data set

Figure 4.11: Bootstrap estimate of the standard error and coefficient of variation for the two verification data sets

The IPS-II scenario

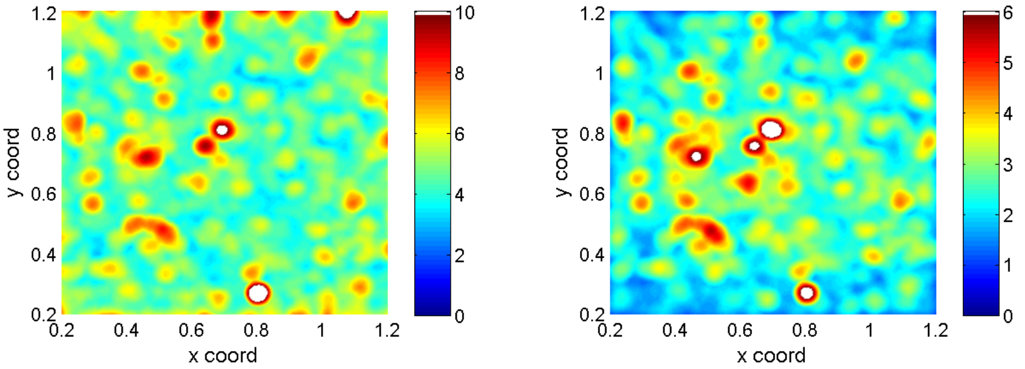
As in the regression analysis, bootstrap estimates of the standard error and coefficient of variation for the plate in the **IPS-II** scenario were produced and presented in Figure 4.13. The average standard error and coefficient of variation obtained from 150 bootstrap re-samples were 2.77 μm and 4.82%, using regression points.



(a) First data set

(b) Second data set

Figure 4.12: Kernel density estimate for the distribution of the (signed) relative error η_s of $\widehat{\sigma}$ for the two new data sets in the **IPS-I** scenario using bootstrap



(a) Bootstrap estimate of the coefficient of variation in %, based on 150 bootstrap re-samples

(b) Bootstrap estimate of the standard error in μm, based on 150 bootstrap re-samples

Figure 4.13: Bootstrap estimate of the standard error and coefficient of variation in the **IPS-II** scenario

The IPS-III scenario

Finally bootstrap was used for estimating the variance of the part of the Volvo, the front wing, considered in Figure 3.15. By applying the above bootstrap method, the resulting estimates of σ presented in Figure 4.14 were obtained.

The average standard error in Figure 4.14 obtained through 100 bootstrap re-samples were about 1 μm. Note that since the actual painting simulation in IPS is so time consuming, this value will not be compared with an ”actual” estimate of the error. However, as bootstrap showed to produce reasonable estimates in the **IPS-I** scenario, there is no

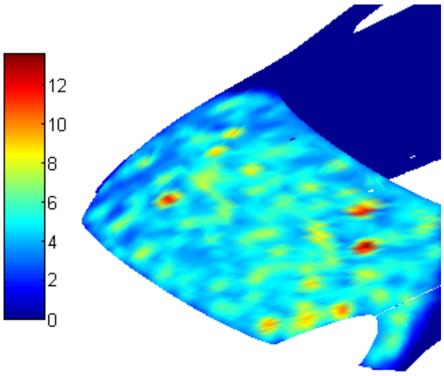


Figure 4.14: Estimated standard error presented in % of the mean estimated thickness for the part of the car under consideration in the **IPS-III** scenario (geometry courtesy of Volvo Car Corporation)

reason to not believe in this result.

4.3 Comparing regression and bootstrap

To easier compare the performance of the different methods the results are summarized and presented next.

S-I scenario

For the **S-I** scenario, all three methods produced estimates of σ within 10% of the value obtained via simulations, see Table 4.5, which summarizes the obtained estimates.

Table 4.5: Estimated $\widehat{\sigma}$ (μm) for the different methods in the **S-I** scenario together with the simulated $\widehat{\sigma}$

Linear regression	Quadratic regression	Bootstrap	Simulations
1.12	1.04	1.18	1.12

IPS-I scenario

Figure 4.15 shows again the distribution of the signed relative error produced through the different methods for the **IPS-I** scenario and in Table 4.6 the corresponding characteristics for the relative error η obtained can be seen. Common for all methods is that they tend produce signed relative errors that are slightly positively skewed [14], meaning that they have heavier tails to the right than to the left.

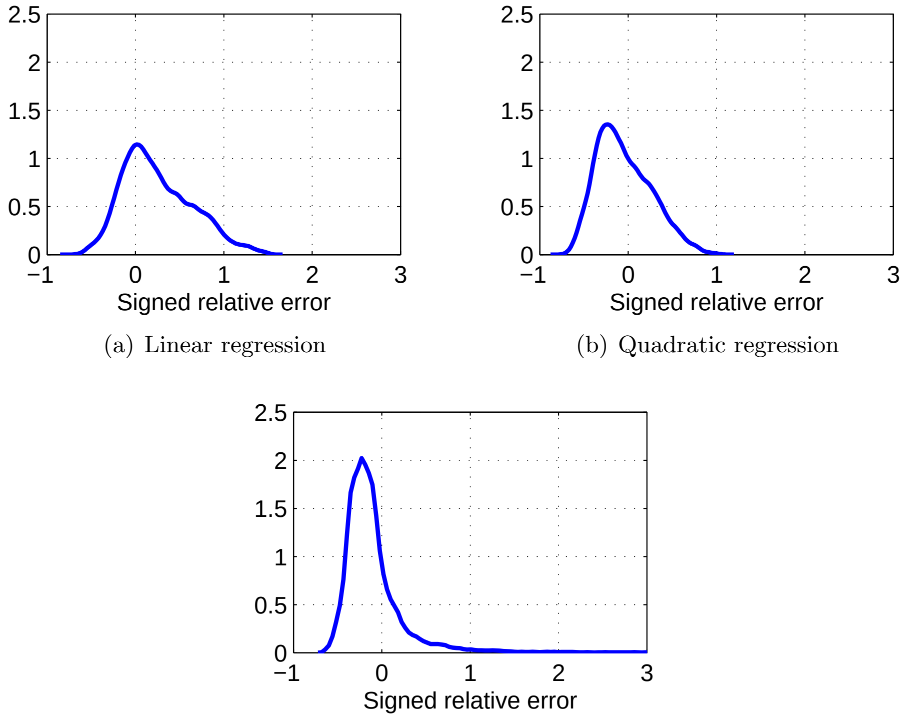
The results obtained through the two verification data sets, given earlier in Table 4.4, shows that the bootstrap produces consistent estimates.

Table 4.6: Relative error for the **IPS-I** scenario using the three different methods

	Linear regression	Quadratic regression	Bootstrap
Mean	0.35	0.26	0.27
Median	0.26	0.24	0.23
Upper 95%	0.98	0.58	0.59

Linear regression yielded larger errors than both bootstrap and quadratic regression. It is however not surprising that the linear model produces less accurate estimates than the quadratic; it does not correct for the nonlinear spatial variations.

Both the original bootstrap estimate together with the two verification data sets did produce some large errors. The maximum values of η obtained were about 3 for the three data sets, which is larger than those obtained from both the regression models.



(a) Linear regression

(b) Quadratic regression

(c) Bootstrap

Figure 4.15: Signed relative error produced through linear and quadratic regression together with the corresponding error obtained from bootstrap for the **IPS-I** scenario

IPS-II scenario

The results obtained in the **IPS-II** scenario can be seen in Table 4.7. Also in this scenario, the methods produced similar estimates.

Table 4.7: Average estimated $\widehat{\sigma}$ (μm) for the **IPS-II** scenario using the different models

	Linear regression	Quadratic regression	Bootstrap
Average $\hat{\sigma}$	2.69	2.24	2.77

5

Summary and discussion

This chapter will discuss the results from the previous chapters, present conclusions and give proposal directions for further work. First comes a short summary giving the main results from the previous chapters.

5.1 Summary

The performance of two different models, denoted the $h-$ and $\lambda-$model, were investigated using synthetic data. Obtained results showed that the $\lambda-$model produced about 10% less error in terms of the square root of the MISE compared to the $h-$model. This method is however more complicated and for simplification, during the rest of the project the $h-$model was used.

Estimates based on anisotropic methods were also investigated. Gradient based methods along with another method based on higher order derivatives, for adjusting the anisotropic methods, were used. Using gradients resulted in better performance than higher order methods.

Two methods for compensating for the lost paint volume when using the kernel density estimation methods for estimating the paint thickness along the boundaries of a surface were used. Although increasing the variance, both methods successfully reduced bias along the boundaries.

Three different methods for estimating the variance of the estimated paint thickness were investigated. Two of them were based on regression models and the third was based on bootstrap to resemble the stochasticity of the painting process. Using bootstrap for producing estimates of the variance of the estimated paint thickness turned out to give values consistent with values obtained from both linear and quadratic regression. The median relative error obtained when estimating σ in the **IPS-I** scenario (one stroke of paint applied to a quadratic plate, see Chapter 2 for a thorough presentation of the different test scenarios) were about 25%, using data from many independent simulations in IPS to obtain more accurate estimates of σ for comparison.

5.2 Discussion and conclusion

Thickness estimation

Since the $\lambda-$model did give about 10% less error in terms of the square root of the MISE than the $h-$model, it is worth to give it a thought whether this model should be used when implementing in IPS (which uses the $h-$model at the moment writing). One drawback with the $\lambda-$ model throughout this thesis was, besides being more complicated, that it required more computational power than the $h-$model. Since `MATLAB` was used, speed was a big problem. When implementing in faster languages, such as `C++`, this might not be much of a problem. Also, instead of using `MATLAB`'s implementation of a kd-tree, faster and more efficient implementations can be used, which would be favorable for the $\lambda-$model.

As expected, a larger value of the cloud factor means that a larger value of the bandwidth should be used for optimizing with respect to the MISE. This is in full agreement with the theory regarding the kernel density estimations presented in the Appendix. A larger cloud factor means a larger variance and increasing the bandwidth will compensate for some of this increased variance with the cost of increasing the error due to bias. From the Tables 3.1 to 3.4 it is clear that using a too large cloud factor, the resulting estimates will be of very low quality. It is also clear that the more spatial variations along the surface of the paint thickness, the smaller cloud factor is needed. This is reasonable, as it would be hard to distinguish noise from actual variations in these cases. As the relationship between optimal bandwidth and cloud factor is different for the different scenarios investigated shows that it is not possible to find a general rule on how to optimally pick a value of the bandwidth based on the cloud factor when considering a new scenario. This is however not the expect since the optimal values depends on the variations in the paint thickness along the surface, according to the theory presented about kernel density estimation given in the Appendix.

It would have been interesting to investigate the **S-II** scenario (synthetic data using the sine function) further by using more values of k. The reason this was not done is because it required a lot of computational effort to find the optimal values, since a large amount of simulations were needed to yield estimates of the MISE.

The anisotropic methods could handle the discontinuity in the intensity of the droplets in the **S-I** scenario (synthetic data using the step function) well by reducing the bias significantly close to the step. As a larger value of h could be used than in the isotropic method, this method also reduces the variance in the estimated thickness for points not close to the step. However, as the support of the kernel was decreased, this lead to higher variance for the estimated thickness in these points. By optimizing the model with respect to the parameters this effect can be reduced to some extent, but there will always be a trade-off between a lower bias and a higher variance.

Anisotropical methods were not enough to correct for the systematically underestimation of the paint thickness along the boundaries of the surfaces where the thickness estimation occurred. The two edge compensation algorithms tested did handle this problem in a better way. In the second method, the thickness was systematically overestimated for points near the boundaries though. A probable reason for this is that by scaling up the volume of a droplet close to the edge, this up-scaling will also effect nearby points lying farther away from the boundary (at a distance large enough to make it unaffected by the original problem). Taking this phenomena into account this method would probably perform better. The first edge compensation algorithm did give satisfying results. What is a bit problematic about the edge compensation algorithms is that it is by no means obvious how to modify them to best fit the general three dimensional surfaces, as for instance the car considered. By linearizing the surface at each point under consideration and then project the nearby surface points on this plane, calculating distances and from this determine if this point is to be considered as being close to the edge (and hence needs to be compensated), one possible method is obtained.

Variance estimation

Bootstrap turned out to give consistent estimates with the regression models for estimating the variance at points where the regression models were applicable. The main difference between the bootstrap and regression models when it comes to estimation of the variance is that bootstrap works by only using the data from which the thickness are estimated. As long as the available methods for performing thickness estimations can be used, bootstrap can be used to yield estimates of the variance. In comparison, the regression models not only need access to the thickness in the specific point where the estimation is made, but also in some nearby points. If the geometry is such that the thickness in those nearby points can not be obtained, then the regression fails.

However, although it has some drawbacks, the regression still has many advantages compared to bootstrap. One of the most important is its speed; it takes much longer to produce estimates through bootstrap.

A possible further development of the variance estimation algorithms could be to combine both bootstrap and regression. Hence, for each re-sampled data set, a thickness is calculated and the variance is estimated by using regression. Repeating this many times and choosing for instance the median estimated variance for each point, or some kind of trimmed mean, a more stable estimate than the one obtained through bootstrap could be obtained.

Bootstrap can be used close to boundaries and at points where the surface has a high curvature. Those can not be handled by the regression models in a satisfactory way. For the bootstrap to give reasonable results it is very important to be able to perform good estimations of the thickness at such points.

The obtained results, presented in Tables 4.5, 4.6 and 4.7 show that bootstrap and the quadratic regression performs similarly well. Though, using bootstrap may produce some very large values of the relative error η in some points. This is because a really heavy droplet in the original data set will make the estimated thickness for the re-sampled data sets sensitive close to impact location of this droplet. If the re-sampling algorithm includes this heavy droplet a number of times (as it will after many runs), the estimated thickness will vary a lot close to it, leading to large values of estimated variance.

5.3 Directions for further work

Next follows some ideas that I find worth to consider for further work in this area:

- Investigate the possible gain by using the $\lambda-$model in IPS instead of the $h-$model. Both when it comes to performance and complexity regarding implementation and computational demands. It could also be interesting to investigate whether this model produces estimates with smaller variance as a result of the more realistic way of dealing with heavy droplets.
- Implement and evaluate either of the edge compensation algorithms for a general 3 dimensional object. Also refined versions of the anisotropic methods are to be tested. Especially, the anisotropic methods both seem to heavily depend on the choice of the parameters used. Therefore a thorough parameter study would be interesting.
- Since bootstrap turned out to give estimates similar to those obtained from the regression models, but with the advantage of being possible to use for the entire object, it should be further investigated and possibly implemented in the existing softwares.
- Although not mentioned in this thesis, the use of importance sampling to reduce the influence of heavy droplets, were initially under consideration. This seems to be a good method for improving the qualities of the estimates. At the moment of writing, this is investigated by other people at FCC, which I highly encourage.

Bibliography

[1] Björn Andersson. Modeling and simulation of rotary bell spray atomizers in automotive paint shops. PhD thesis, Chalmers University of Technology, Gothenburg, 2013.

[2] Robert Andersson and Christoffer Johansson. Paint thickness estimation: A poisson model of droplet impacts. Bachelor's thesis, Chalmers University of Technology, Gothenburg, 2011.

[3] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[4] Tarn Duong. Bandwidth Selectors for multivariate kernel density estimation. PhD thesis, University of Western Australia, 2004.

[5] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 3rd edition, 2001.

[6] Michael T Heath. *Scientific Computing: an Introduction Survey*. McGraw-Hill New York, New York, 2nd edition, 2002.

[7] Wolfgang Härdle, Marlene Müller, and Stefan Sperlich. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer Verlag Berlin Heidelberg, 1st edition, 2004.

[8] John Isaksson. Statistical Methods for Estimating Variance of Paint Thickness on Curved Surfaces. Master's thesis, Chalmers University of Technology, Gothenburg, 2013.

[9] Hemant M Kakde. Range Searching using Kd Tree. [Accessed 7th May 2014] `http://www.cs.utah.edu/~lifeifei/cs6931/kdtree.pdf`, 2005.

[10] Otto JWF Kardaun. *Classical Methods of Statistics*. Springer Verlag Berlin Heidelberg, Garching, Germany, 2005.

[11] Håkon Kile. Bandwidth Selection in Kernel Density Estimation. Master's thesis, Norwegian University of Science and Technology, May 2010.

[12] David C Lay. *Linear Algebra and its Applications.* Addison-Wesley, New York, 4th edition, 2011.

[13] Zhi Ouyang. Univariate Kernel Density Estimation. [Accessed 7th May 2014] `http://www.stat.duke.edu/~zo2/shared/research/readings/kernelsmoothing.pdf`, 2005.

[14] John A Rice. *Mathematical statistics and data analysis.* Duxbury Advanced Series. Brooks/Cole, 3rd edition, 2007.

[15] Igor Rychlik and Jesper Rydén. *Probability and risk analysis: an introduction for engineers.* Springer, 2006.

[16] Lars Schjøth, Jon Sporring, and Ole Fogh Olsen. Diffusion based photon mapping. In *Computer Graphics Forum*, volume 27, pages 2114–2127. Wiley Online Library, 2008.

[17] Roy L Streit. *Poisson Point Processes, Imaging, Tracking, and Sensing*, volume 1. Springer, Reston, Virginia, 2010.

[18] S Tafuri, F Ekstedt, J Carlson, A S Mark, and F Edelvik. Improved spray paint thickness calculation from simulated droplets using density estimation. *Proceedings of the ASME 2012 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE*, 2:339–347.

[19] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Society, Vienna, Austria, 2012.

[20] Sanford Weisberg. *Applied linear regression*, volume 528 of *Wiley Series in Probability and Statistics.* John Wiley & Sons, 3rd edition, 2005.

[21] Walter Zucchini. Part 1: Kernel density estimation. In *Applied smoothing techniques.* Philadelphia, Pa. Temple University, 2003.

A

Appendix

Here follows an introduction to the mathematical and statistical concepts that will be used throughout the thesis.

A.1 Some probability theory and statistics

Inverse sampling

To sample from a random variable X with invertible cumulative distribution function F, let U be a sample from a uniform [0,1] variable and let $Y = F^{-1}(U)$. Then since $\mathbf{P}\left(Y \leq x\right) = \mathbf{P}\left(F^{-1}(U) \leq x\right) = \mathbf{P}\left(U \leq F(x)\right) = F(x)$, the sample Y is distributed according to $F(x)$ [14].

Sampling from nonstandard probability distributions

When generating samples from a probability distribution where the inverse sampling method cannot easily be used, the acceptance-rejection sampling method is an alternative method [3]. Suppose we want to generate samples from a continuous density $f(x)$. Let $h(x)$ be another density from which we can easily generate samples, with the property that for some constant c, $f(x) \leq ch(x)\ \forall x$. Generate a sample Z from h and a sample U from the uniform distribution on [0,1] and calculate the ratio $f(Z)/ch(Z)$. If this ratio is greater than U, then we accept Z as a realization of f, but if it is less we reject it.

The performance of the acceptance-rejection sampling algorithm depends mainly on how tight f can be dominated by ch, the larger the c the larger the number of rejections and hence the slower the algorithm.

Bias of an estimator

The bias of an estimator $\hat{\theta}$ of θ is a systematic error and defined as $\mathbf{E}[\hat{\theta}] - \theta$ [20]. An estimator is said to be unbiased if $\mathbf{E}[\hat{\theta}] - \theta = 0$, *i.e.* if it is correct ”on average”.

Variance

Variance is a measure of how much a random variable X differs from its mean on average and is mathematically defined as [14]

$$\mathrm{Var}\,(X) = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2.$$

Estimation of the variance, often denoted σ^2, of a random variable X is usually done by drawing n independent samples, $X_1, X_2, ...X_n$ from X and calculating the sample variance [14]

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2, \tag{A.1}$$

which is an unbiased estimate of the variance σ^2, *i.e.* $\mathbf{E}s^2 = \sigma^2$. Estimating the mean $\mu = \mathbf{E}X$ of a random variable X is done by approximation with the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ which is an unbiased estimate of μ [14]. The standard error for $\bar{X}$ is $s_{\bar{X}} = s/\sqrt{n}$ for i.i.d. sampling of sample size n.

The normalized standard deviation,

$$c_v = \frac{\sigma}{\mu},$$

is called *Coefficient of Variation* and is used to quantify the variations in relation to the mean [15].

Bootstrap

Bootstrap techniques are computationally intensive Monte Carlo methods which can be used for estimating sampling distributions when the setup is too complex to allow for analytical methods [14]. Suppose we want to find an estimate of $\theta = g(X_1, X_2, ..., X_n)$ where g is some function from $\mathcal{R}^n \to \mathcal{R}$ and $X_1, X_2, ..., X_n$ are independent and identically distributed random variables and the distribution of $g(X_1, X_2, ..., X_n)$ is not known. Under some technical details which we assume hold, by the Law of Large Numbers [5], picking n realizations of X, calculating $g(X_1, X_2, ..., X_n)$ and repeating the procedure m times and averaging gives an estimate

$$\widehat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(X_1, X_2, ..., X_n),$$

of θ which converges in probability as $m \to \infty$. However, it is not always possible or desirable to generate new random variables, and here bootstrap techniques come handy.

Given n independent realizations $x_1, x_2, ..., x_n$ of the random variable X, an estimate of θ is obtained by calculating $g(X_1, X_2, ..., X_n)$. This will however only result in a point estimate and will not give any further information about the accuracy of the estimate, the error or its distribution. If n is large, the empirical cdf [14] obtained through the samples can be used to approximate the distribution of X, hence by drawing n samples from $x_1, x_2, ..., x_n$ with replacement we get a new data set $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ from which we can calculate estimates of θ. This is often repeated thousands or even more times to get a good estimate of the sampling distribution of θ, from which the accuracy of the estimate can be obtained.

Regression analysis

Regression analysis can be used to fit models to data in a way that minimizes a certain error, often in the least square sense [14]. Next follows an explanation of multiple regression analysis of linear and quadratic models.

Linear models

Linear models are of the form [14]

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1} + e,$$

where Y is the response variable, β_k is an unknown parameter to be estimated, x_k is the value of the kth variable (called predictor variable), from which we want to explain the variations in Y that depends on the variations of the k'th variable and e is a random noise (also called error), assumed to be $\mathcal{N}(0, \sigma^2)$ distributed.

For n given sets of predictor variables $x_{i,1}, x_{i,2}, ..., x_{i,p-1}$, $i = 1, 2, ..., n$ and n realizations $y_i, ..., y_n$ of the response variable Y, the task is to find the parameters $\beta_0, ..., \beta_{p-1}$ in a way that minimizes the errors. In matrix notation the linear regression model is written $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{\beta} = (\beta_0, ..., \beta_n)^T$, $\boldsymbol{e} = (e_1, ..., e_n)^T$ and

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & \dots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{pmatrix},$$

where $\boldsymbol{X}$ is called the design matrix. Hence, we wish to find $\widehat{\boldsymbol{\beta}}$ such that [14] it minimizes

$$S(\widehat{\boldsymbol{\beta}}) = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2.$$

$\widehat{\boldsymbol{\beta}}$ can be calculated through the normal equations [20] as

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}.$$

An unbiased estimate of the common variance σ^2 of the errors e can be obtained via [14]

$$s^2 = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n-p}.$$

Nonlinear models

For the linear model to perform well, the response variable has to linearly depend on the predictor variables. If this is not the case, the regression model has to be corrected. Linear regression can be generalized to also handle polynomial models [20] of which the quadratic model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + e,$$

with two predictor variables is a special case. This model is in fact also linear in the parameters β_i and hence the previous theory regarding linear regression also applies to the quadratic model, and especially the estimate

$$s^2 = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n-p},$$

is still an unbiased estimate of σ^2 [14].

The lognormal distribution

X is said to be lognormally distributed, $X \sim \log \mathcal{N}(\mu, \sigma^2)$ with parameters μ and σ, if the logarithm of X is normally distributed, $\log X \sim \mathcal{N}(\mu, \sigma^2)$, with parameters μ and σ [5]. Figure A.1 shows the probability density function for some combinations of μ and σ.

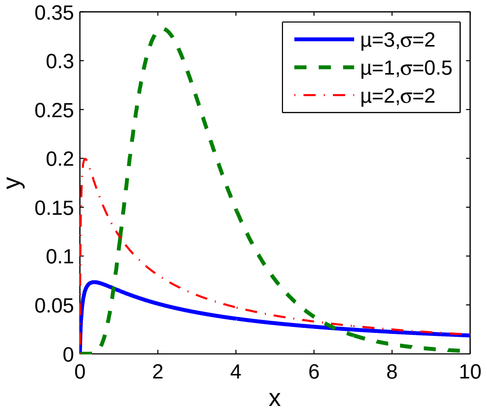


Figure A.1: The lognormal probability density function for some combinations of the parameters μ and σ

A.2 Nonparametric density estimation

There are various methods for estimating probability density functions, both parametric, semi parametric and nonparametric [7]. Here we will focus on nonparametric methods. The easiest and most commonly used method among these is the histogram [7].

Histogram

Suppose that $X_1, X_2, ..., X_n$ are n realizations of a random variable X and assume further that they are independent. Using these samples we would like to estimate the probability density function f_X without imposing any assumptions about the distribution beforehand. First partition a part of the real line covering the samples $X_1, X_2, ..., X_n$ into bins, b_j with a width of h. The fraction of observations falling into each bin is then calculated, *i.e.*

$$\widehat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n}\sum_{j}\mathbf{1}(X_i \in b_j)\mathbf{1}(x \in b_j),$$

yielding a piecewise constant function $\widehat{f}_X$. It can be shown [7] that this is a consistent estimator of the probability density function, meaning that by letting the number of observations $n \to \infty$ while letting the binwidth $h \to 0$ the function $\widehat{f}_X$ will converge in probability to f_X. The histogram is however not an unbiased estimator. In Figure A.2 a histogram for 10 000 independent $\mathcal{N}(5,1)$ variables can be seen together with the true (weighted) density.

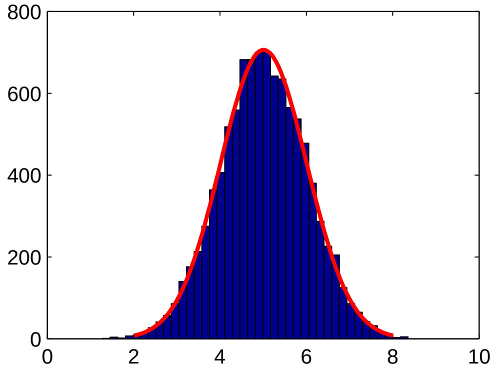


Figure A.2: Histogram together with the density for a $\mathcal{N}(5,1)$ variable for 10 000 random samples

Although being easy to use and interpret, there are some undesirable drawbacks using the histogram as an estimate for the probability density function. It is very sensitive to the placement and size of the bins [7, 13] and using a step function as an estimate of a continuous function is in many cases not satisfactory [21]. This leads us to kernel density estimates.

Kernel density estimates

Kernel density estimation is a method resembling the histogram method but without the major drawbacks [7]. Let X be a random variable with pdf $f(x)$. Then [21]

$$\mathbf{P}(X \in (x-h, x+h)) = \int_{[x-h,x+h]} f(z)\,\mathrm{d}z \approx 2hf(x),$$

so hence

$$f(x) \approx \frac{1}{2h}\mathbf{P}\left(X \in (x-h, x+h)\right).$$

The probability can be estimated using the relative frequency in the sample, yielding the estimate

$$\widehat{f}(x) \approx \frac{\#\text{ observations in the interval } (x-h, x+h)}{2hn} = \frac{1}{n}\sum_{i=1}^{n} w(x-x_i, h),$$

where w is a weighting function such that

$$w(x) = \begin{cases} 1/2h & \text{if } x \in (x-h, x+h) \\ 0 & \text{if } x \notin (x-h, x+h) \end{cases}.$$

The kernel

More generally, a kernel density estimate can be written as

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h}),$$

where K is a kernel function [21] [18] with the properties that

$$\begin{aligned} \int_{\mathbf{R}} K(x)\,\mathrm{d}x = 1, \\ K(x) = K(-x), \\ \int_{\mathbf{R}} x^2K(x)\,\mathrm{d}x = m_2 < \infty. \end{aligned} \tag{A.2}$$

There are various of different kernel functions [21] and in Table A.1 some one dimensional kernel functions can be seen.

In Figure A.3 a kernel estimate of a standard normal probability density function, together with the corresponding estimate produced through the histogram method, based on a set of independent $\mathcal{N}(0,1)$ realizations. Note the much smoother behavior of the kernel density estimator.

Table A.1: One dimensional kernel functions

Kernel	$K(x)$
Gaussian	$\frac{1}{2\pi}e^{-x^2/2}$
Epanechnikov	$\frac{3}{4}\left(1-x^2\right)\mathbf{1}_{\lvert x\rvert<1}$
Triangular	$(1-\lvert t\rvert)\,\mathbf{1}_{\lvert x\rvert<1}$
Rectangular	$\frac{1}{2}\mathbf{1}_{\lvert x\rvert<1}$

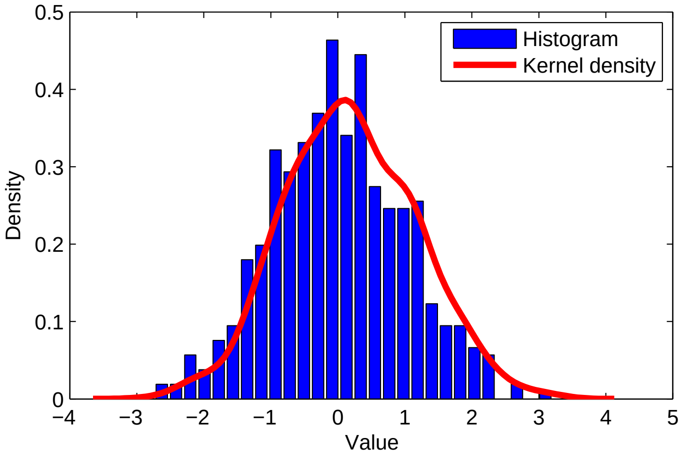


Figure A.3: Histogram together with the kernel density estimate for a $N(0,1)$ variable

Bandwidth

There is a trade-off between the variance and the bias of the estimate. For larger values of h the variance of the estimate will decrease while the bias will increase and for smaller values of h the variance increases while the bias decreases [18].

A common way to quantify the accuracy of an estimate is the mean squared error, MSE, defined as [4] [21]

$$\begin{aligned}
\mathrm{MSE}(\widehat{f}(x)) &= \mathbf{E}\left[\widehat{f}(x) - f(x)\right]^2 \\
&= \mathbf{E}\widehat{f}^2(x) - 2f(x)\mathbf{E}\widehat{f}(x) + f^2(x) \\
&= \left(\mathbf{E}\widehat{f}(x)\right)^2 + f^2(x) - 2f(x)\mathbf{E}\widehat{f}(x) + \mathbf{E}\widehat{f}^2(x) - \left(\mathbf{E}\widehat{f}(x)\right)^2 \qquad \text{(A.3)} \\
&= \left(\mathbf{E}\widehat{f}(x) - f(x)\right)^2 + \mathbf{E}\left[\widehat{f}^2(x)\right] - \left(\mathbf{E}\widehat{f}(x)\right)^2 \\
&= \mathrm{Bias}^2(\widehat{f}(x)) + \mathrm{Var}\widehat{f}(x).
\end{aligned}$$

The bias and the variance of the kernel estimate satisfies [7]

$$\mathrm{Bias}\widehat{f}(x) = \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2), \text{ as } h \to 0, \qquad \text{(A.4)}$$

and

$$\mathrm{Var}\widehat{f}(x) = \frac{1}{nh}\|K\|_2^2 f(x) + o\left(\frac{1}{nh}\right), \text{ as } nh \to \infty, \qquad \text{(A.5)}$$

where $\mu_2(K) = \int_{\mathbf{R}} x^2 K(x)\,\mathrm{d}x$ and $\|K\|_2^2 = \int_{\mathbf{R}} K^2(x)dx$, *i.e.* the squared L^2 norm. Clearly, the choice of bandwidth h is crucial for the performance of the kernel density estimation.

Combining Equation A.3 with the Equations A.4 and A.5 the MSE of the kernel density estimate can be written as [7]

$$\mathrm{MSE}\widehat{f}(x) = \frac{h^4}{4}f''(x)^2\mu_2(K)^2 + \frac{1}{nh}\|K\|_2^2 f(x) + o(h^4) + o(\frac{1}{nh}). \qquad \text{(A.6)}$$

The MSE in Equation A.6 goes to zero as $h \to 0$ and $nh \to \infty$. The MSE is however a local measure of estimation accuracy as it depends on x. A global measure is the mean integrated squared error [7], MISE, defined as

$$\begin{aligned}
\mathrm{MISE}\widehat{f} &= \mathbf{E}\left[\int_{\mathbf{R}} (\widehat{f}(x) - f(x))^2\,\mathrm{d}x\right] \\
&= \int_{\mathbf{R}} \mathbf{E}\left[(\widehat{f}(x) - f(x))^2\right]\mathrm{d}x \qquad \text{(A.7)} \\
&= \int_{\mathbf{R}} \mathrm{MSE}\widehat{f}(x)\,\mathrm{d}x.
\end{aligned}$$

Equation A.6 together with Equation A.7 yields

$$\mathrm{MISE}\widehat{f} = \frac{h^4}{4}\|f''(x)\|^2\mu_2(K)^2 + \frac{1}{nh}\|K\|_2^2 + o(h^4) + o(\frac{1}{nh}),$$

and ignoring terms of higher order we arrive at the asymptotic mean integrated squared error, AMISE

$$\mathrm{AMISE}\widehat{f} = \frac{h^4}{4}\|f''(x)\|^2\mu_2(K)^2 + \frac{1}{nh}\|K\|_2^2.$$

Multivariate kernel density estimation

Just as in the one dimensional case mentioned above, when estimating multivariate d dimensional density functions, multivariate kernel density estimation is a nonparametric approach. Let $\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_n}, \boldsymbol{X_j} \in \mathbf{R}^d$ be n realizations of the d dimensional random variable $\boldsymbol{X}$ with probability density function $f(\boldsymbol{x})$. An estimate of $\boldsymbol{x}$ can then be written as [7]

$$\begin{aligned}\widehat{f}_h(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\prod_{j=1}^{d} h_j}\mathcal{K}(\frac{\boldsymbol{x}-\boldsymbol{X_i}}{\boldsymbol{h}}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\prod_{j=1}^{d} h_j}\mathcal{K}(\frac{x-X_{i1}}{h_1}, \frac{x-X_{i2}}{h_2}, ..., \frac{x-X_{id}}{h_d}),\end{aligned}$$

where $\mathcal{K} : \mathbf{R}^d \to \mathbf{R}$ is a kernel function with d arguments satisfying similar properties as in Equation A.2 as the kernel function K in the univariate case and $\boldsymbol{h} = (h_1, h_2, ...h_d) \in \mathbf{R}^d$ is a vector of bandwidths.

Two different ways of constructing multivariate kernel functions from univariate kernel functions are mentioned in [7], the multiplicative kernel

$$\mathcal{K}(\boldsymbol{x}) = \prod_{i=1}^{d} K_i(x_i),$$

and the radial symmetric kernel

$$\mathcal{K}(\boldsymbol{x}) = \alpha K(\|\boldsymbol{x}\|),$$

where α is a normalization constant to make sure the kernel integrates to 1. For the two dimensional Epanechnikov kernel, the normalization constant is as follows:

$$\begin{aligned}\int_{\mathbf{R}^2}\mathcal{K}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int_{x^2+y^2\leq 1}(1-(x^2+y^2))\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^{2\pi}\int_0^1 (1-r^2)r\,\mathrm{d}r\,\mathrm{d}\phi \\ &= 2\pi\left[\frac{2r^2-r^4}{4}\right]_0^1 \\ &= \pi/2.\end{aligned}$$

Hence, $\alpha = 2/\pi$ for the two dimensional Epanechnikov kernel.

As in the univariate case, the choice of bandwidths are crucial. A more general approach is to use a bandwidth matrix $\mathbf{H} \in \mathcal{R}^{d\times d}$, which is symmetric, positive definite and nonrandom [4]. This will make it possible to anisotropically adapt the bandwidth [18]. The estimate now take the general form

$$\widehat{f}(\boldsymbol{x};\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\det\mathbf{H}}\mathcal{K}\left(\mathbf{H}^{-1}(\boldsymbol{x}-\boldsymbol{X_i})\right).$$

According to [7] the bias and variance are

$$\mathrm{Bias}\widehat{f}(\boldsymbol{x};\mathbf{H}) \approx \frac{1}{2}\mu_2(\mathcal{K})\mathrm{tr}(\mathbf{H}^T D^2 f(\boldsymbol{x})\mathbf{H}),$$

and

$$\mathrm{Var}\widehat{f}(\boldsymbol{x};\mathbf{H}) \approx \frac{1}{n\det\mathbf{H}}\|\mathcal{K}\|_2^2 f(\boldsymbol{x}),$$

where tr(.) is the trace of a matrix and $D^2 f(\boldsymbol{x})$ denotes the Hessian matrix of f [12]. Hence the MSE of the multivariate kernel density estimate is

$$\mathrm{MSE}\widehat{f}(x;\mathbf{H}) \approx \frac{1}{4}\mu_2^2(\mathcal{K})\mathrm{tr}(\mathbf{H}^T D^2 f(\boldsymbol{x})\mathbf{H})^2 + \frac{1}{n\det\mathbf{H}}\|\mathcal{K}\|_2^2 f(\boldsymbol{x}).$$

Integrating yields an estimate of the MISE

$$\mathrm{MISE}\widehat{f}(x;\mathbf{H}) \approx \frac{1}{4}\mu_2^2(\mathcal{K})\int_{\mathcal{R}^d}\mathrm{tr}(\mathbf{H}^T D^2 f(\boldsymbol{x})\mathbf{H})^2\,\mathrm{d}\boldsymbol{x} + \frac{1}{n\det\mathbf{H}}\|\mathcal{K}\|_2^2. \tag{A.8}$$

Changing the approximate equality to an equality in the equation above, we arrive at the AMISE

$$\mathrm{AMISE}\widehat{f}(x;H) = \frac{1}{4}\mu_2^2(\mathcal{K})\int_{\mathcal{R}^d}\mathrm{tr}(\mathbf{H}^T D^2 f(\boldsymbol{x})\mathbf{H})^2\,\mathrm{d}\boldsymbol{x} + \frac{1}{n\det\mathbf{H}}\mathcal{K}\|_2^2.$$

We want to minimize the MISE with respect to the choice of bandwidth, *i.e.* we want to find [4]

$$\mathbf{H}_{\mathrm{MISE}} = \underset{\mathbf{H}\in\mathcal{H}}{\mathrm{argmin}}\ \mathrm{MISE}\widehat{f}(.,\mathbf{H}).$$