Creating a Pseudo-CT from MRI for MRI-only based Radiation Therapy Planning

Daniel Andreasen

Kongens Lyngby 2013
IMM-MSc-2013-10

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk IMM-MSc-2013-10

Summary (English)

Background: In the planning process of external radiation therapy, CT is used as the main imaging modality. The advantage of using CT is that the voxel intensity values are directly related to electron density which is needed for dose calculations. Furthermore, CT provides an accurate geometrical representation of bone needed for constructing digitally reconstructed radiographs. In recent years, interest in replacing CT with MRI in the treatment planning process has emerged. This is due to the fact that MRI provides a superior soft tissue contrast; a desirable property that could increase the accuracy of target and risk volume delineation. The challenge in replacing CT with MRI is that the MRI intensity values are not related to electron densities and conventional MRI sequences cannot obtain signal from bone.

The purpose of this project was to investigate the use of Gaussian Mixture Regression (GMR) and Random Forest regression (RaFR) for creating a pseudo-CT image from MRI images. Creating a pseudo-CT from MRI would eliminate the need for a real CT scan and thus facilitate an MRI-only work flow in the radiation therapy planning process. The use of GMR for pseudo-CT creation has previously been reported so the reproducibility of these results was investigated. dUTE and mDixon MRI image sets as well as Local Binary Pattern (LBP) feature images were investigated as input to the regression models.

Materials and methods: Head scans of three patients fixated for whole brain radiation therapy were acquired on a 1 T open MRI scanner with flex coils. dUTE and mDixon image sets were obtained. CT head scans were also acquired using a standard protocol. A registration of the CT and MRI image sets was carried out and LBP feature images were derived from the dUTE image sets.

All RaFR and GMR models were trained with the dUTE image sets as basic input. Some of the models were trained with an additional mDixon or LBP input in order to investigate if these inputs could improve the quality of the predicted pseudo-CT. More specifically, the impact of adding the LBP input was investigated using RaFR and the impact of adding an mDixon input was investigated using both RaFR and GMR. A study of the optimal tree depth for RaFR was also carried out. The quality of the resulting pseudo-CTs was quantified in terms of the prediction deviation, the geometrical accuracy of bone and the dosimetric accuracy.

Results: In the LBP input study, the results indicated that using LBPs could improve the quality of the pseudo-CT.

In the mDixon input study, the results suggested that both RaFR and GMR models were improved when adding the mDixon input. The improvement was mainly observed in terms of smaller prediction deviations in the bone region of the pseudo-CTs and a greater geometrical accuracy. When comparing RaFR and GMR, it was found that using RaFR produced pseudo-CTs with the smallest prediction deviations and greatest geometrical accuracy. In terms of the dosimetric accuracy, the difference was less clear.

Conclusion: The use of GMR and RaFR for creating a pseudo-CT image from MRI images was investigated. The reproducibility of previously reported results using GMR was demonstrated. Furthermore, the impact of adding LBP and mDixon inputs to the regression models was demonstrated and showed that an improvement of the pseudo-CT could be obtained. The results serves as a motivation for further studies using more data and improved feature images.

Summary (Danish)

Baggrund: I planlægningen af ekstern stråleterapi anvendes CT som den primære skanningsmodalitet. Fordelen ved at anvende CT er, at voxel-intensitetsværdier er direkte relateret til elektrontætheden af vævet, der afbildes. Dette er nødvendigt for at kunne udføre dosisberegninger. Desuden giver CT en nøjagtig geometrisk repræsentation af knogle, som er nødvendig for at generere digitalt rekonstruerede røntgenbilleder afledt fra CT. I de seneste år er interessen for at erstatte CT med MRI i planlægningsprocessen opstået. Dette skyldes det faktum, at MRI er CT overlegen, når det kommer til bløddelskontrast; en vigtig egenskab, der kan øge indtegningsnøjagtigheden af tumorvolumener og risikoorganer. Udfordringen i at erstatte CT med MRI er, at MRI intensitetsværdier ikke er direkte relateret til elektrontæthed samt at konventionelle MRI sekvenser ikke kan opnå signal fra knogle.

Formålet med dette projekt var at undersøge brugen af Gaussian Mixture Regression (GMR) og Random Forest regression (RaFR) for at skabe et pseudo-CT billede ud fra MRI-billeder. Dannelsen af et pseudo-CT fra MRI kan potentielt eliminere behovet for en CT-scanning og dermed muliggøre et work flow baseret udelukkende på MRI. Anvendelsen af GMR til at generere pseudo-CTer er tidligere blevet rapporteret, så et mål med projektet var at undersøge reproducerbarheden af resultaterne. dUTE og mDixon MRI skanninger samt Local Binary Pattern (LBP) featurebilleder blev undersøgt som input til regressionsmodellerne.

Materialer og metoder: Tre patienter som skulle modtage helhjerne strålebehandling blev skannet på en 1 T åben MRI-skanner med flex-spoler. dUTE og mDixon billedsæt blev optaget. En CT skanning af hovedet blev også optaget

ved hjælp af standardprotokol. CT- og MRI-billedsæt blev indbyrdes registreret og LBP featurebilleder blev afledt fra dUTE billedsættene. Alle RaFR og GMR modeller blev trænet med dUTE billedsæt som grundlæggende input. Nogle af modellerne blev trænet med et yderligere mDixon eller LBP input med henblik på at undersøge om disse billedsæt kunne forbedre kvaliteten af pseudo-CTerne. Specifikt blev virkningen af at tilføje LBP input undersøgt for RaFR og virkningen af at tilføje mDixon input undersøgt for både RaFR og GMR. En undersøgelse af den optimale *tree depth* for RaFR blev også udført. Kvaliteten af de resulterende pseudo-CTs blev kvantificeret ved hjælp af den gennemsnitlige afvigelse mellem pseudo-CT og reference CT, den geometriske nøjagtighed af knogle i pseudo-CTet samt den dosimetriske nøjagtighed når pseudo-CTet blev brugt til at beregne dosefordelinger.

Resultater: I undersøgelsen af LBP input viste resultaterne, at anvendelsen af LBP kunne forbedre kvaliteten af pseudo-CTet.

I undersøgelsen af mDixon input viste resultaterne, at både RaFR og GMR modellerne blev forbedret ved brug af mDixon input. Forbedringerne gav sig hovedsagligt til udtryk i mindre afvigelser i knogleregionen af pseudo-CTerne samt ved større geometrisk nøjagtighed. Ved sammenligning af RaFR og GMR fandtes det, at pseudo-CTer genereret med RaFR havde de mindste gennemsnitlige afvigelser samt største geometriske nøjagtighed. Med hensyn til den dosimetriske nøjagtighed var forskellen mellem GMR og RaFR ikke tydelig.

Konklusion: Anvendelsen af GMR og RaFR for at generere pseudo-CT-billeder fra MRI-billeder blev undersøgt. Reproducerbarheden af tidligere rapporterede resultater ved anvendelse af GMR blev demonstreret. Endvidere blev virkningen af at tilføje LBP og mDixon input til regressionsmodellerne undersøgt. Det blev vist, at en forbedring af pseudo-CTet derved kunne opnås. Resultaterne tjener som en motivation for at udføre yderligere undersøgelser med mere data og forbedrede featurebilleder.

Preface

This thesis was prepared at the department of Informatics and Mathematical Modelling at the Technical University of Denmark and at the department of Oncology at Copenhagen University Hospital, Herlev. The work corresponds to 30 ECTS points and it was done in partial fulfilment of the requirements for acquiring an M.Sc. in Medicine and Technology at the Technical University of Denmark and the University of Copenhagen.

Supervisors
Jens M. Edmund, Ph.D., DABR
Department of Oncology (R)
Copenhagen University Hospital, Herlev

Koen Van Leemput, Ph.D.
Department of Informatics and Mathematical Modelling
Technical University of Denmark

Lyngby, 01-March-2013

Daniel Andreasen

Acknowledgements

First of all, I would like to thank the Department of Oncology at Herlev Hospital for making this project possible and allowing me to work in an exciting clinical environment. It has given me great insight and I have felt very privileged. I would like to thank Rasmus H. Hansen, who set up the MRI scanner for the special sequences used in my project, and Jon Andersen, who was in charge of recruiting patients. Without their help, much of this project would not have been possible.

Also, a great thanks to both of my supervisors, Jens M. Edmund and Koen Van Leemput for their guidance, enthusiasm and great help with this project.

Finally, I would like to thank my friends and family for their support and feedback throughout the project. A special thanks to my girlfriend, Minna, whose support and faith in me has been invaluable in stressful times.

Acronyms and symbols

Acronyms

CART	Classification and Regression Tree
CNR	Contrast-to-Noise Ratio
CT	Computed Tomography
DRR	Digitally Reconstructed Radiograph
dUTE	difference UTE (dual echo UTE)
DVH	Dose-Volume Histogram
EM	Expectation Maximization
FOV	Field of View
GMM	Gaussian Mixture Model
GMR	Gaussian Mixture Regression
GTV	Gross Tumour Volume
Gy	gray
HU	Hounsfield unit
ICRU	The International Commission on Radiation Units & Measurements
LBP	Local Binary Pattern
LINAC	Linear accelerator

MAPD	Mean Absolute Prediction Deviation
mDixon	multi-echo Dixon
MPD	Mean Prediction Deviation
MRI	Magnetic Resonance Imaging
MSPD	Mean Squared Prediction Deviation
MU	Monitor Unit - a standardized measure of machine output for a LINAC
OR	Organ at Risk
pCT	pseudo-CT
pdf	Probability Density Function
PET	Positron Emission Tomography
PRV	Planning Risk Volume
PTV	Planning Target Volume
RaF	Random Forest
RaFR	Random Forest Regression
RF	Radio Frequency
RT	Radiation Therapy
SNR	Signal-to-Noise Ratio
std	Standard Deviation
T	Tesla
UTE	Ultra Short Echo Time

List of symbols

α	Flip angle
β	Split parameters of a node in a decision tree
dT_E	Time interval between echo acquisitions in the multi-echo Dixon sequence
F	Magnitude of the magnetization of fat
ϕ	Multivariate Gaussian Probability density function
π_j	Prior weight in a Gaussian Mixture Model
T_1	Longitudinal magnetization relaxation time constant
T_2	Transversal magnetization relaxation time constant
T_2^*	Time constant of the free induction decay
T_{AQ}	Acquisition time
T_E	Echo time
Θ	Phase angle
θ	Parameters of a Gaussian Mixture Model
φ	Error phase due to field inhomogeneities
φ_0	Error phase due to MRI system flaws
W	Magnitude of the magnetization of water

Contents

Chapter 1

Introduction and Background

Despite the growing interest in using magnetic resonance imaging (MRI) in the planning process of external radiation therapy (RT), Computed Tomography (CT) remains the golden standard. This is due to the fact that a planning CT scan provides electron density information which is critical to the calculation of the 3D dose distribution in the irradiated tissues. Furthermore, the CT provides an accurate geometrical representation of bone, which is needed for constructing a digitally reconstructed radiograph (DRR); a 2D planar reference image created from the CT image set. This is used in combination with traditional X-ray images taken at the linear accelerator (LINAC) to verify proper patient setup with respect to the isocenter of the LINAC [1].

In conventional external RT planning, CT is the only imaging modality used. A consequence of this is that delineation of e.g. the tumour volume and potential organs at risk (ORs) must be done on the CT image sets. In 2006, Khoo and Joon outlined the advantages of introducing MRI in the planning process [2]. Their main point was that MRI contributes with an increased soft tissue contrast, yielding a better characterization of tissues, which in CT appear very similar. Kristensen *et al.* later showed that delineating structures on MRI images leads to more accurate volume definitions [3]. For this reason, an MRI/CT work flow has become common practice at Herlev Hospital where structures are

delineated on MRI image sets and then transferred to the CT image sets for dose plan calculation. The transfer of delineations is carried out using a rigid registration between the MRI and CT image sets.

Aside from requiring an extra scan of the patient, the MRI/CT work flow has some disadvantages. Nyholm *et al.* estimated that systematic spatial uncertainties of 2 mm are introduced with the rigid registration between MRI and CT [4]. The term *systematic* refers to the fact that the same error is repeated over and over, as opposed to the random errors introduced with e.g. a less than perfect patient alignment with respect to the LINAC. The systematic errors may in other words cause a constant partial miss of the tumour volume and potentially result in an increased dose to organs at risk (ORs).

1.1 MRI-only Radiation Therapy Planning

With the advances of MRI in RT treatment planning and with the introduction of PET/MRI systems, the interest and research in entirely replacing CT with MRI has increased [5–7]. This would eliminate the registration induced systematic errors and reduce the amount of scanning sessions needed. The approach has been to generate an estimate of CT images from MRI images, a so called pseudo-CT (pCT) which could then be used for dose calculation in RT treatment planning and for attenuation correction in PET/MRI. The challenge is that there is no direct relation between MRI intensity and electron density. Furthermore, conventional MRI cannot depict cortical bone very well because the bone signal is lost before the readout begins. This is due to the short T_2 of cortical bone.

1.1.1 dUTE Imaging

In 2003, Robson *et al.* described the basic ultra short echo time (UTE) MRI sequence and its ability to acquire signal from bone and ligaments [8]. The so-called difference UTE (dUTE) sequence was also introduced as a means of enhancing the visibility of short T_2 components. This was achieved by acquiring a gradient echo shortly after the first acquisition and then subtracting the two image sets to isolate the short T_2 components. Several authors have later used the dUTE sequence for pCT estimation or attenuation correction [5,9–11]. At Herlev Hospital, Kjer investigated the dUTE sequence and found optimal acquisition parameters for the 1 Tesla (T) open MRI scanner on site [12]. With

regard to these parameters, it should be noted that at the stated optimal second echo time of the dUTE, phase cancellation artefacts may occur due to the chemical shift between water and fat. This in turn may cause non-bone tissue to appear as bone (see Section 2.1 for more details).

1.1.2 Dixon Imaging

In the field of PET/MRI some of the same challenges are present as in MRI-only RT treatment planning. Here, attenuation maps for attenuation correction need to be derived. This was previously obtained from CT using its direct correlation with electron density. Because it has been found important to account for adipose tissue in whole-body MRI-based attenuation correction, the Dixon MRI sequence has been investigated for its possible use in generating attenuation maps [9, 13, 14]. This sequence provides water-only and fat-only image sets and for this reason is appropriate for segmenting water and fat in MRI images.

1.1.3 Estimating CT from MRI

Different approaches have been used to generate pCTs. One is an atlas-based estimation. With this method, an atlas of many co-registered MRI and CT image sets is made. To estimate a pCT for a new patient, the MRI atlas can be deformed to fit the new patient's MRI. The same deformation can then be applied to the CT atlas to obtain an estimate of the pCT. A variation of the standard atlas-based method was done by Hofmann *et al.* [15]. They used patches in T_1 weighted images and the spatial position from an atlas as features to train a regression model used to predict the pCT. The fact that no special MRI sequences, such as the dUTE, was needed to generate the pCT seems to be the main strength of this atlas-based method. In general however, atlas-based methods are known to suffer from inaccuracies when the inter-patient variability is large [16].

Another approach has been to do a voxel-wise segmentation of the MRI. Here a characterization of each voxel in the MRI is done in order estimate a pCT. A common characteristic of the voxel-wise methods seems to be the need for special MRI sequences like the dUTE or Dixon in order to compensate for the previously mentioned shortcomings of conventional MRI.
In the field of PET/MRI Martinez-Möller *et al.* used thresholds to segment Dixon MRI images into air/lung/fat/soft tissue classes which were then assigned

bulk attenuation factors. They did not take bone into account [13]. Berker *et al.* used a combined dUTE/Dixon sequence to obtain an air/bone/fat/soft tissue segmentation, again using thresholds [9]. Both Martinez-Möller *et al.* and Berker *et al.* reported improvements in segmentation accuracy when using the Dixon sequence. They quantified the accuracy of their methods by looking at the agreement with CT attenuation correction and not at the similarity between the pCT and real CT, which makes it hard to compare their method with others'.

Instead of segmenting the MRI images into different tissue classes, Johansson *et al.* tried to predict the CT image by using a Gaussian mixture regression (GMR) method. Data was collected from 5 patients who were CT scanned (one image set) and MRI scanned on a 1.5 T scanner with two dUTE sequences at different flip angles and one T_2 weighted sequence (5 image sets in total). Each of the 5 MRI image sets were filtered using a mean filter and a standard deviation (std) filter to create 10 more image sets (5 mean filtered and 5 std filtered). An air mask was used to reduce the data amount and the regression model was then estimated using k-means and expectation maximization (EM) algorithms to estimate the mixing proportions, mean values and covariance matrices. The pCT was compared to the actual CT image using the mean absolute prediction deviation (MAPD) inside the air mask. It was found to be 137 HU on average for all five patients with a variation from 117 to 176 HU for individual patients. The method showed very promising results and was able to distinguish bone from air. The largest differences between pCT and CT were seen at bone/tissue and air/tissue interfaces (paranasal sinuses). This was attributed to susceptibility effects causing a very short T_2^*. In a later publication, Johansson *et al.* explored the voxel-wise uncertainty of the model and used the mean prediction deviation (MPD) and MAPD to determine if model reduction was possible [16]. Here, they found that the T_2 weighted image set was redundant in their model.

1.2 Random Forest Regression

Random Forest regression (RaFR) has to the knowledge of the author not been used for predicting pCTs from MRI prior to this project. However, it has been used to predict organ bounding boxes in both MRI Dixon and CT image sets with promising results [17,18]. To predict organ bounding boxes in Dixon MRI image sets, Pauly *et al.* used a local binary pattern-like feature in order to describe each pixel by its neighbourhood appearance [17]. The regression task was a bit different from that of predicting a pCT since they were not trying to predict a CT value but a distance to a bounding box for every MRI voxel.

1.3 Objectives

Inspired by the approach of Johansson *et al.*, a regression approach was taken on in this project in order to find a mapping of MRI intensity values into CT densities. More specifically, the Gaussian mixture regression model was used in order to test its robustness and the reproducibility of results in a different clinical set up. Motivated by the promising results of Pauly *et al.* Random Forest regression was studied as a method for predicting pCTs. Also, the use of Local Binary Pattern (LBP) as input to RaFR was studied.

Building on the prior experiences with voxel-wise methods, it was expected that special MRI sequences would be needed to generate a pCT. For this reason the dUTE acquisition method was adopted. The data was collected using the scanner-specific dUTE acquisition parameters reported by Kjer. From the model reduction study by Johansson *et al.*, the method of using two different flip angles for the dUTE and adding filtered images to the regression models showed to improve the prediction accuracy. These methods were also adopted.

As mentioned, phase cancellation artefacts in water/fat containing voxels were suspected to be present in the dUTE image sets; an issue specific for the 1 T scanner at Herlev Hospital. Based on the experience that the Dixon sequence can help in distinguishing water and fat it was investigated whether the sequence could improve the regression models used.

With this established, the main objectives of the project can be summarized as follows:

1. Investigate the use of Gaussian mixture regression for pCT generation.
2. Investigate the use of RaFR for pCT generation. This includes looking into optimizing the model parameters and using a Local Binary Pattern-like feature.
3. Investigate the impact of using a Dixon sequence in both models.
4. Compare the performance of Gaussian mixture regression versus that of Random Forest regression for pCT generation.
5. Quantify the differences between the reference CT and the pCT in measures relevant for radiation therapy.

Chapter 2

Theory

2.1 dUTE MRI Sequence

Conventional MRI sequences are able to begin the readout of signal at a minimum T_E of about 8-10 ms [8]. This makes them inappropriate for detection of signal from tissues that lose their transversal magnetization faster than this. Tissues such as cortical bone, periosteum and ligaments which have a T_2 in the range 0.05-10 ms thus appear with a similar intensity as air in conventional MRI images.

The ultra short echo time (UTE) MRI pulse sequence is optimized to detect signal from tissues with a short T_2. This means that an unconventional acquisition approach has been taken in order to minimize the time from excitation to readout and to maximize the amount of signal coming from short T_2 components in this time frame. Because the image acquired with a UTE sequence shows high signal intensities from both tissues with a short and a long T_2, it is common to record a second (or dual) echo shortly after the first. This technique is referred to as the difference UTE (dUTE) technique. In the resultant second image set the tissues with short T_2 will have lost a significant amount of signal compared to the long T_2 tissues. The second image can thus be subtracted from the first to identify/isolate the short T_2 components (see Figure 2.1).

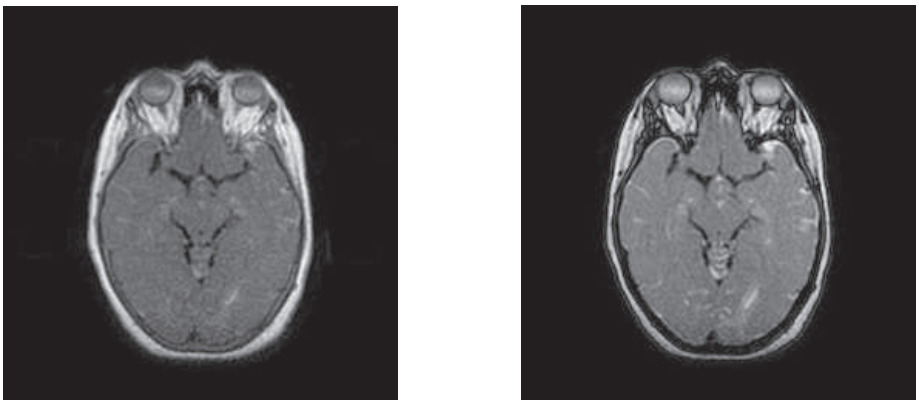


Figure 2.1: Dual echo UTE images. **Left:** Image acquired at $T_E = 0.09\ ms$. **Right:** Image acquired at $T_E = 3.5\ ms$. As can be seen signal has been lost from short T_2 components in the second echo image.

Below, the UTE deviations from conventional MRI sequences will be outlined. It should be noted that these parameter deviations are related to the acquisition of the *first* echo of the dUTE sequence, since this is recorded with an ultra short echo time. The second echo is a conventional gradient echo.

2.1.1 Parameters

In general, 3 parameters are important when acquiring signal from short T_2 components. These are the RF pulse duration, the echo time T_E and the acquisition time T_{AQ}.

RF pulse duration. In conventional MRI imaging the duration of the excitation RF pulse is not a concern since the T_2 of the tissues being imaged is longer than the pulse duration. However, when dealing with tissues with a short T_2, loss of signal during the RF pulse becomes a problem [19]. To maximize the transversal component of the magnetization in short T_2 components, short RF pulses are used which ensures the least amount of T_2 relaxation during excitation. A consequence of this is that the flip angle becomes lower than the conventional 90° (typically 40-70° less). It is important to note, that this lower flip angle actually produces *more* signal from short T_2 components, contrary to what one might intuitively think.
In a 3D UTE sequence a single RF pulse of short duration is used, where after 3D radial readout gradients are used to traverse k-space.

Echo time. To get the most amount of signal from short T_2 components the optimal T_E would ideally be 0 ms. This is not possible because the MRI coils used for both excitation and signal acquisition need a little time to switch from transmit to receive mode. This is a physical limitation that depends on the scanner hardware used and in practice the shortest possible T_E is used.

Acquisition time. With conventional MRI sequences k-space is traversed in a linear rectangular manner. To save time, sampling in UTE imaging is done in a non-linear fashion and simultaneous with the ramp up of the readout gradient, which leads to a radial (or centre-out) sampling. This in turn means that k-space becomes oversampled close to the centre and thus low spatial frequencies have a higher signal-to-noise ratio (SNR) than high ones [19]. T_{AQ} is the sampling duration and this must be short enough for the short T_2 tissues to not lose signal before the end of acquisition. On the other hand, some time is needed to traverse a certain distance from the centre of k-space in order to capture high spatial frequency components. In practice, a compromise must be made that maximizes signal and minimizes blurring, which means a T_{AQ} of approximately the T_2 of the tissue being imaged [19].

2.1.2 dUTE Artefacts

Because of the ultra short echo time, it is actually the free induction decay that is measured during the readout of the first echo (which is thus not an echo in the traditional sense). This means that it is impossible to distinguish between relaxation due to T_2 effects and relaxation due to T_2^* effects. However, for fast relaxing components, it is reasonable to assume that $T_2 \approx T_2^*$ [20]. This may not hold at tissue interfaces where susceptibility effects cause de-phasing within voxels due to field inhomogeneities. This yields signal intensity artefacts because tissue with an otherwise long T_2 loses signal rapidly due to a short T_2^* induced by the field inhomogeneity.

Concerning the second echo acquisition, another artefact worth noting is the chemical shift or phase cancellation artefact. Because hydrogen bound in water has a slightly different resonance frequency than that of hydrogen in fat, the signal from these will at certain times after excitation periodically be in or out of phase. At out of phase times, less or no signal will be present in voxels containing a mixture of water and fat. Since the phase cancellation is time dependent, the severity of the artefact depends on the chosen T_E. The chemical shift (or difference in resonance frequencies) is measured in parts per million (ppm) and for water and fat it is 3.4 ppm. At 1 T, water has a resonance frequency of 42 MHz, which in turn means that water and fat will be in phase at a frequency of 3.4 ppm $\cdot$ 42 MHz = 142.8 Hz. This corresponds to every

7 ms. The first time after excitation when water and fat are out of phase is thus 3.5 ms. As mentioned, Kjer previously investigated the optimal acquisition parameters for the dUTE sequence at the MRI scanner at Herlev Hospital. He found that a T_E of 3.5 ms was close to the optimal echo time for the second echo in the dUTE sequence in terms of contrast-to-noise ratio (CNR) of the dUTE image sets [12]. The dUTE image sets recorded on the 1 T scanner are thus susceptible to phase cancellation artefacts.

2.2 Dixon MRI Sequence

The Dixon technique was invented by W.T. Dixon in 1984 [21]. In short, the technique facilitates separating the signal originating from water and fat in MRI images, which makes it possible to make water-only and fat-only images. The rationale is that, in many standard MRI sequences, fat appears hyper-intense, thus covering signal from other contents. This means that water-containing structures of interest or contrast agents may be hard to discern in high-fat regions [22]. Also, the chemical shift artefact between water and fat may cause phase cancellation or spatial shifts which is undesirable.
In its original form, the Dixon technique is relatively simple. It is based on the recording of two images, one where water and fat are in phase ($\Theta = 0°$) and one were they are 180° out-of-phase ($\Theta = 180°$), a so-called 2-point method (because of the number of images recorded). By representing the MRI data as a complex signal it can be described using its magnitude and phase. If it is assumed that the imaged object only consists of water and fat, the signal can be described as [23]:

$$S(x,y) = [W(x,y) + F(x,y) \cdot e^{i\Theta}] \cdot e^{i\varphi(x,y)} \cdot e^{i\varphi_0(x,y)}$$

where x and y are spatial coordinates, W and F are the magnitudes of the magnetizations of water and fat, Θ is the phase angle between water and fat, φ is an error phase due to field inhomogeneities and φ_0 is an error phase due to various minor system flaws. Disregarding the field inhomogeneity and the spatial coordinates the two signals acquired at $\Theta = 0°$ and $\Theta = 180°$ can be written as [23]:

$$S_0 = (W + F) \cdot e^{i\varphi_0} \tag{2.1}$$

$$S_{180} = (W - F) \cdot e^{i\varphi_0} \tag{2.2}$$

Which can be rewritten to:

$$W = \frac{|S_0 + S_{180}|}{2} \tag{2.3}$$

$$F = \frac{|S_0 - S_{180}|}{2} \tag{2.4}$$

This is the basic principle behind Dixon's technique. In reality, however, field inhomogeneities *are* present (along with noise and artefacts) and equations 2.3 and 2.4 are not sufficient to provide a good separation. Consequently there is a need to estimate φ in order to use the Dixon technique for water/fat separation. It turns out that this estimation is not straightforward. If ϕ is taken into account, Equation 2.2 becomes [23]:

$$S_{180} = (W - F) \cdot e^{i\varphi} e^{i\varphi_0} \tag{2.5}$$

and ϕ can then be estimated as:

$$\hat{\varphi} = \frac{\arctan\left((S_{180} \cdot S_0^*)^2\right)}{2} \tag{2.6}$$

The challenge here lies in the fact that the arctan operation limits the phase estimation to be in the interval $[-\pi, \pi]$, i.e. the phase is wrapped. If the actual phase is above or below the aforementioned interval a multiple of 2π will be added or subtracted to the estimated value to make it fit in the interval. The consequence of phase wrapping is that water voxels may be misinterpreted as fat voxels and vice versa. Different approaches have been taken to compensate for the phase wrapping using both 1-point, 2-point or 3-point recordings and with different phase recovering algorithms, but it is beyond the scope of this report to go in to details with these.

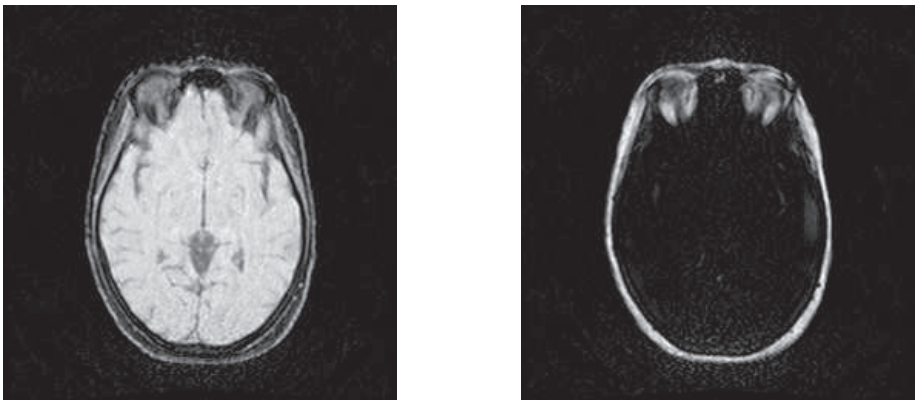


Figure 2.2: Dixon water and fat separated images. **Left:** Image with a majority of water containing components. **Right:** Image with a majority of fat containing components.

As with most MRI sequences, the scan time of the Dixon sequence should be minimized to reduce motion artefacts and patient discomfort. Most of the Dixon sequences require the acquisition of more than one image and as such they are more susceptible to motion artefacts. A way to reduce scan time is by using multiple readouts per excitation, the so called multi-echo Dixon (mDixon) technique [24]. In this way data can be sampled for both images in one excitation increasing the efficiency. See Figure 2.2 for an example of Dixon water and fat images.

2.3 Registration Method

The CT and MRI images are recorded on two different scanning facilities and as a result, these image sets are often of different resolution and spatially unaligned. Training a regression model on unaligned data makes little sense and the need to ensure that one voxel in the MRI corresponds to the same voxel in the CT is thus evident. Below follows a description of the alignment method used in this project.

A *rigid registration* involves positioning the MRI and CT in the same frame of reference and then translate and/or rotate one of the images until the desired correspondence with the other (stationary) image is achieved. An assumption is that little or no anatomical differences exist from one image to the other. This is often assumed for an intra-patient registration of scans done at approximately the same instance in time.

The mutual information method provides a tool for registering multi-modality image sets [25]. To arrive at the definition of mutual information, a description of the joint histogram of grey-values is needed.

The joint histogram is a 2D representation of the combination of grey-values for all corresponding voxels in two image sets. The appearance of the joint histogram is thus determined by the alignment of the two image sets. If they are perfectly registered, there is an overlap between all air voxels in the two image sets. This will show as a clearly defined cluster in the joint histogram because many air-air voxel combinations are counted. Similar clusters will appear for overlapping tissue voxels. If the two images are not aligned, the joint histogram will conversely show a large dispersion because many different combinations of grey-values appear.

The entropy can measure the dispersion of a probability distribution. By normalizing the joint histogram with the total amount of observations, the joint probability density function (pdf) of grey-values can be estimated. The entropy of the joint probability density function (pdf) is now defined as [26]:

$$H(A,B) = -\sum_{a,b} pdf(a,b) \log pdf(a,b) \tag{2.7}$$

where A is the image set being transformed, B is the stationary image set and $pdf(a,b)$ is the joint probability of observing the overlapping grey-values a and b in image A and B, respectively. The joint entropy is calculated for the overlapping region of the two image sets and by measuring the dispersion of the joint pdf it can be used to determine the degree of alignment between the image sets. A low entropy means a low degree of dispersion.

Equation 2.7 provides a means of registering images but it is not always robust. One may encounter situations where the overlapping regions show a low entropy even though they are not spatially correlated. To avoid this, the marginal entropies of grey-values in image set A and B (in the overlapping region) can be taken into account. This yields the mutual information, I [26]:

$$I(A,B) = H(A) + H(B) - H(A,B) \tag{2.8}$$

where the marginal entropies are calculated as

$$H(A) = -\sum_a pdf(a) \log pdf(a)$$

and

$$H(B) = -\sum_b pdf(b) \log pdf(b)$$

with the marginal densities calculated from the joint pdf as $pdf(a) = \sum_b pdf(a,b)$ and $pdf(b) = \sum_a pdf(a,b)$.

The registration strategy then involves finding the transformation parameters which maximize the mutual information. Typically a crude alignment of the image sets is done first, for instance by aligning the centroids, then an optimization algorithm is used to maximize the mutual information.

As the transformation parameters are continuous in nature the transformed image coordinates will almost inevitably fall in between voxel coordinates, which means an interpolation of the transformed image to the new voxel coordinates is necessary.

2.4 Gaussian Mixture Regression

This section describes the theory behind regression using a Gaussian mixture model (GMM).

2.4.1 Gaussian Mixture Model

The foundation for doing Gaussian mixture regression (GMR) is to model the joint density of the input/output (X/Y) space by a weighted sum of K multivariate Gaussian probability density functions (pdfs) [27]:

$$f_{X,Y}(x,y) = \sum_{j=1}^{K} \pi_j \phi(x,y;\mu_j,\Sigma_j) \tag{2.9}$$

here, π_j is a prior weight subject to the constraint $\sum_{j=1}^{K} \pi_j = 1$ and ϕ is a multivariate Gaussian pdf with mean $\mu_j = \begin{bmatrix} \mu_{jX} \\ \mu_{jY} \end{bmatrix}$ and covariance matrix $\Sigma_j = \begin{bmatrix} \Sigma_{jXX} & \Sigma_{jXY} \\ \Sigma_{jYX} & \Sigma_{jYY} \end{bmatrix}$. By definition Σ_j is symmetric so $\Sigma_{jXY} = \Sigma_{jYX}$. Equation 2.9 is called a Gaussian mixture model (GMM). Each Gaussian, or *component*, in the model is sought to explain the distribution of a sub-population in the data.

In the context of this project, the input, X, consists the MRI image sets and the output, Y, is the corresponding CT image set. x and y are spatially corresponding voxel intensity values in the MRI and CT image sets, respectively. Equation 2.9 is then used to model the joint distribution of voxel intensities in the image sets. The assumption is that, even though there are variations in the appearance of the same tissue type from one MRI scan to another, the intensities will still centre around a mean value. This in turn means that, even though there is no direct relation between an MRI and a CT intensity value, their joint distribution will still consist of a discrete amount of tissue clusters [5].

The parameters $\theta_j = (\pi_j, \mu_j, \Sigma_j)$ of each Gaussian in the Gaussian mixture model (GMM) are often not known in advance and need to be estimated from the data at hand. A common way of doing this is by maximizing the log likelihood function which explains the probability of the data given the parameters [28]:

$$\hat{\theta} = \underset{\theta}{argmax}(p(\boldsymbol{X}, \boldsymbol{Y}|\theta)) \tag{2.10}$$

where $\boldsymbol{X}, \boldsymbol{Y}$ denotes the data. The expectation maximization (EM) algorithm can be used to achieve this. This optimization method iteratively estimates the maximum likelihood parameters from an initial guess of the parameter values. It is beyond the scope of this report to go into the details on how the expectation maximization (EM) algorithm optimizes the log likelihood, but two things should be noted about the method. Firstly, the EM algorithm cannot determine the number of components to use. This means that, for a good estimation of the GMM, one needs a prior knowledge as to the number of components or subpopulations that exist in the data. Furthermore, the EM method may converge to a local (and not global) maximum of the log likelihood function depending on the initial starting point. Hence, the initial parameter guess is quite important as it may affect which optimum is found. In order to come up with a qualified initial guess on the composition of the components in the GMM, a k-means clustering algorithm can be applied to the data to make a rough estimation of $\hat{\theta}$. This does not solve all problems because the k-means clustering algorithm also needs to be provided with initial values of the centres of the clusters (so-called *seeds*). In this project, the seeds were chosen as K random samples from the training data. This has the consequence that the result from the k-means clustering (and thus the initial parameters for the EM algorithm) may vary each time a model is trained even if the same training data is used.

2.4.2 Regression using a Gaussian Mixture Model

Gaussian mixture regression (GMR) consists of a training phase and a test phase. When a decision has been made as to the number of components to use in the GMM, the training phase is, as described in the previous sections, composed of estimating the parameters of the GMM using the EM algorithm. Once the GMM has been estimated, it can be used for regression. This is the test phase, which means the GMM is used on previously unseen input data, to make a prediction on the appearance of the output. In relation to this project, the input test data would be the MRI image sets of a new patient and the predicted output would be the pseudo-CT (pCT). To make predictions, the expected value of Y given an observed value of $X = x$ should be found. From the definition of joint density, each Gaussian in Equation 2.9 can be partitioned as the product of the conditional density of Y and the marginal density of X of each Gaussian:

$$\phi(x, y; \mu_j, \Sigma_j) = \phi(y|x; m_j(x), \sigma_j^2)\phi(x; \mu_{jX}, \Sigma_{jXX}), \ j \in 1, 2, ..., K \tag{2.11}$$

where

$$m_j(x) = \mu_{jY} + \Sigma_{jYX}\Sigma_{jXX}^{-1}(x - \mu_{jX}) \tag{2.12}$$

$$\sigma_j^2 = \Sigma_{jYY} - \Sigma_{jYX}\Sigma_{jXX}^{-1}\Sigma_{jXY} \tag{2.13}$$

are the conditional mean function and variance of Y. Inserting the result of Equation 2.11 into Equation 2.9 yields:

$$f_{X,Y}(x,y) = \sum_{j=1}^{K} \pi_j \phi(y|x; m_j(x), \sigma_j^2) \phi(x; \mu_{jX}, \Sigma_{jXX}) \tag{2.14}$$

The conditional density of Y|X is now defined as:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{2.15}$$

Where

$$f_X(x) = \int f_{Y,X}(y,x)\,\mathrm{d}y = \sum_{j=1}^{K} \pi_j \phi(x; \mu_{jX}, \Sigma_{jXX}) \tag{2.16}$$

is the marginal density of X. Inserting the definitions of $f_{X,Y}(x,y)$ and $f_X(x)$ into Equation 2.15 finally yields:

$$f_{Y|X}(y|x) = \sum_{j=1}^{K} \frac{\pi_j \phi(y; m_j(x), \sigma_j^2) \phi(x; \mu_{jX}, \Sigma_{jXX})}{\sum_{j=1}^{K} \pi_j \phi(x; \mu_{jX}, \Sigma_{jXX})} \tag{2.17}$$

This can also be expressed as:

$$f_{Y|X}(y|x) = \sum_{j=1}^{K} w_j(x) \phi(y; m_j(x), \sigma_j^2) \tag{2.18}$$

with the mixing weight:

$$w_j(x) = \frac{\pi_j \phi(x; \mu_{jX}, \Sigma_{jXX})}{\sum_{j=1}^{K} \pi_j \phi(x; \mu_{jX}, \Sigma_{jXX})} \tag{2.19}$$

The expected value of Y for a given $X = x$ can now be found as the conditional mean function of Equation 2.18:

$$E[Y|X = x] = m(x) = \sum_{j=1}^{K} w_j(x) m_j(x) \tag{2.20}$$

Equation 2.20 is the regression function in a GMR model and as can be seen, once the GMM has been established all the parameters needed for regression is actually contained in Equation 2.9. In other words, the crucial part of doing GMR lies in estimating a proper GMM. A simple example of GMR illustrating the steps involved can be seen in Figure 2.3.

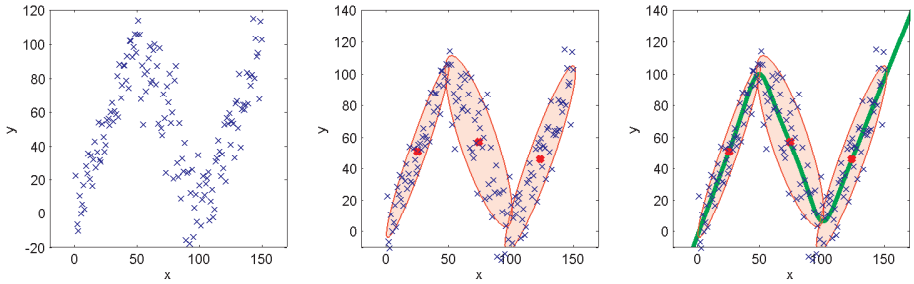


Figure 2.3: Illustration of GMR using simple univariate input and output. **Left:** Data generated by adding Gaussian noise to three linear functions. **Middle:** A GMM consisting of $K = 3$ components is estimated using the EM algorithm with k-means initialisation, mean values are marked as red dots. **Right:** GMR estimates the expected value of y given 400 values of x in the interval $[-10, 170]$, marked in green.

2.4.3 Impact of Changing the Number of Components

Once the k-means algorithm is used with random seeds to make the initial parameter guess for the EM algorithm, the only parameter to tweak in GMR is the number of components to use in the GMM. As mentioned, ideally, one should have a prior knowledge as to the number of sub-populations in the data in order to estimate a model that explains the data well. In the example in Figure 2.3 it was known a priori that the regression problem consisted of three linear functions and for this reason it made sense to use three components to model it. In Figure 2.4, the impact of varying the number of components is illustrated. As can be seen, using too many components leads to over-fitting the data, which is undesirable if noise is present. On the other hand, using too few components leads to a higher degree of smoothing of the data, which may mean that important variations are missed. With real-world multidimensional data it can be quite hard to know how many components should be used. A way to find the optimal number of components is to set up a measure of regression accuracy, evaluate different models and choose the one that performs the best. The mean squared prediction deviation (MSPD) is a measure which has been used before [29]. To be able to evaluate the performance, one needs to have a ground truth to compare the predictions against. This means that training must be done on one part of the data at hand and testing on another. A model is then estimated using the training data which is then used to predict the response on the test data. The mean squared error can then be calculated and used to quantify the quality of the model. Formally the mean squared prediction

deviation (MSPD) can be written as:

$$MSPD = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N} \tag{2.21}$$

where N is the data size, y_i is the i'th true output value of the test data and $\hat{y}_i$ is the i'th predicted value when applying the regression model on the test data.

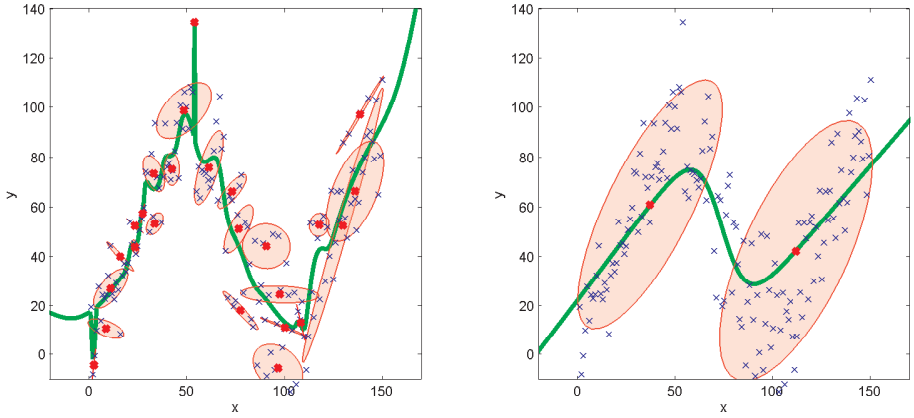


Figure 2.4: Illustration of GMR using a different number of components. Data is the same as in Figure 2.3. **Left:** The GMM has been estimated using 25 components. **Right:** The GMM has been estimated using 2 components.

2.5 Random Forest Regression

The concept of Random Forests (RaFs) is as an extension to the classification and regression tree (CART) methodology introduced by Breiman *et al.* in 1984 [30]. It basically consists of using an ensemble of weak classifiers or predictors for classification or regression. Using an ensemble shows to improve the accuracy of classification or prediction compared to using a single classifier or predictor [31].

In this section the use of a Random Forest (RaF) for regression will be described and illustrated. Examples in this section have been generated by our own RaF implementation based on a paper by Criminisi *et al.* [32]. It should be noted, however, that the method used for producing the pCTs later in this report was based on Breiman's RaF as implemented in the Statistics toolbox of Matlab R2012a [31]. Our own implementation provided greater parameter flexibility than the Matlab version, and hence, was a good tool for illustrating simple RaFR. It did, however, prove to be inferior in robustness when handling large amounts of data which is why it was not used for the actual pCT predictions. Some minor differences exist between the method by Breiman (who uses the CART methodology for tree training) and the one by Criminisi *et al.*. These will also be outlined below.

2.5.1 Decision Trees

The decision tree is the basic building block of a random forest. The main principle consists of pushing data through so-called *nodes* and at each node do a binary test on the data in order to split it in two. The process is repeated on the split data to do another split and this continues until a stop criterion is met. When sketched, the resulting structure of the nodes resembles that of a tree and, to keep with this analogy, the last nodes of the tree are similarly called *leaf* nodes, see Figure 2.5. When using decision trees for regression (called a regression tree), a prediction is made at each leaf node as to what the value of the output variable should be given the input data that reached the node.

2.5.1.1 Constructing a Regression Tree

In order to create a regression tree, training data is used to estimate the optimal split parameters of the binary decision functions at each node. Criminisi *et al.* uses a continuous formulation of information gain as a measure of the split

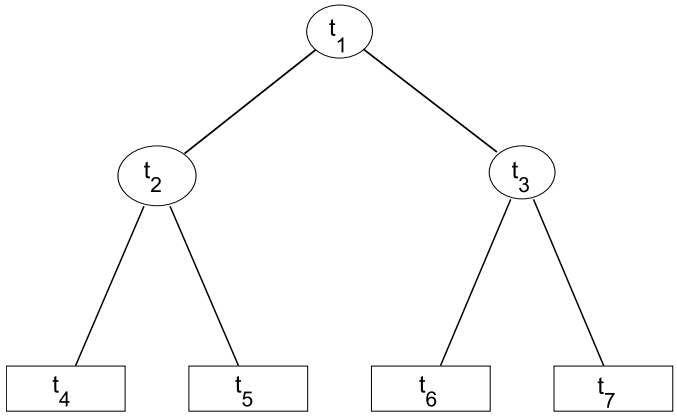


Figure 2.5: Illustration of a regression tree. t_1 is the root node. t_2-t_3 are the internal nodes of the tree. t_4-t_7 are the so-called *leaf* nodes.

quality:

$$I_j = \sum_{x_j \in S_j} \log(\sigma_y(x_j)) - \sum_{i \in \{L,R\}} \left(\sum_{x_j \in S_j^i} \log(\sigma_y(x_j)) \right) \tag{2.22}$$

where x and y is the input and output training data, S_j denotes the set of data reaching node j, σ_y is the conditional variance of y given x_j found from a linear least squares model fit to y_j. $S_j^i, i \in L, R$ denotes the set of training data reaching the left (L) and right (R) node below node j. To put it simply, by maximizing Equation 2.22, one finds the split which minimizes the uncertainty at all training points of a linear fit. Breiman's measure of split quality is the least absolute deviation:

$$LAD(d) = \frac{1}{N_j} \sum_{x_j \in S_j} |y(x_j) - d(x_j)| \tag{2.23}$$

where N_j is the size of the data at node j, $d(x_j)$ is the predicted value of the output at input x_j and $y(x_j)$ is the actual value. It turns out that the predictor that minimizes the LAD is the median of $y(x_j)$, denoted $\upsilon(y(x_j)))$. To put it in a similar form as Equation 2.22, the LAD measure for a split of node j into it's left and right nodes can be written as:

$$I_j = \sum_{x_j \in S_j} |y(x_j) - \upsilon(y(x_j))| - \sum_{x_j \in S_j^L} |y(x_j) - \upsilon(y(x_j))| - \sum_{x_j \in S_j^R} |y(x_j) - \upsilon(y(x_j))| \tag{2.24}$$

Maximizing Equation 2.24 corresponds to finding the optimal split of node j by minimizing the sum of the absolute deviations from the left and right node medians.

Now that a measure of the optimal split has been established, one needs to decide the structure of the binary decision function. The function takes the form:

$$h(\boldsymbol{u}, \beta_j) \in \{0, 1\} \tag{2.25}$$

where $\boldsymbol{u}$ is a vector of the input variables (or features) and β_j are the split parameters at node j. The split is achieved by choosing one or more of the input variables and then choosing a threshold on the value of those variables (or a linear combination thereof) to decide which data goes left and right.

The training of the regression tree is now straightforward. At each node, Equation 2.22 or 2.24 is maximized with respect to β_j to find the thresholds and features that best split the data. This can be done either by an exhaustive search or by a specialized update algorithm [31]. The found parameters are saved at each node. Growing of the tree stops when a pre-specified minimum amount of data reaches a node, when all output training data that reaches a node has the same value or the tree reaches a pre-specified depth (number of node levels). At the leaf nodes Criminisi *et al.* save the parameters of a least squares linear regression model fitted to the data at each node. Breiman saves the median of the output training data at each leaf.

2.5.1.2 Prediction

Once the tree has been constructed, making predictions on test data is done by sending the data through the nodes and splitting it using the learned split parameters. Once the leaf nodes are reached, a prediction can be made. Using Breiman's method the predicted value is the stored leaf node medians whereas Criminisi *et al.* apply the linear regression model to the data. In other words, using Breiman's method the prediction is a constant at each leaf node whereas it is a function of the input test data in Criminisi *et al.*'s method.

2.5.2 Random Forests

Random Forest regression (RaFR) is an ensemble method. It consists of training a number of regression trees, each randomly different from each other, and then using the average of all tree outputs as the forest output. Each tree should be a weak predictor of the output, which means that they should not be fully optimized as described in Section 2.5.1.1. Instead randomness is induced when training each tree to ensure de-correlation between the outputs of each tree. This can be done by either training each tree on a random subset of the training data, called *bagging* (short for *bootstrap aggregating*), or by only letting a

random subset of split parameters be available at each node, called random node optimization. A combination of the two can also be used. In Figure 2.6, a simple case of regression using a random forest is visualized. In the left panel, artificial data has been generated by adding Gaussian noise to three linear functions. In the middle panel, the predictions of 200 randomly different regression trees are shown. In the right panel, the final prediction of the random forest is shown. A maximum tree depth of 3 was used, meaning that the largest tree could consist of a root node, two internal nodes and 4 leaf nodes. If less than 5 data points reached a node it was declared a leaf node. Random node optimization was used, meaning that at each node 10 randomly chosen splits where made available.

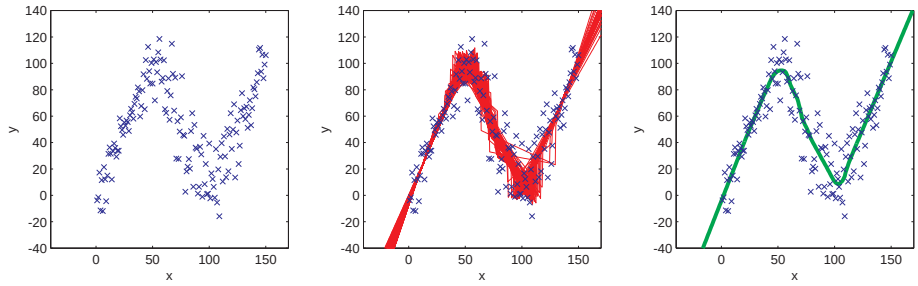


Figure 2.6: Illustration of regression using a random forest. **Left:** Artificial data. **Middle:** Predictions of 200 random regression trees shown in red. **Right:** The final prediction of the random forest shown in green.

2.5.3 Effect of Forest Size

In general, the larger an ensemble of trees, the better predictions. The prediction accuracy will, however, converge towards a constant as the number of trees are increased. Adding more trees hereafter, will only increase the prediction accuracy very little. An advantage of random forests is that increasing the number of trees does not lead to over-fitting. It will, however, increase the computation time and in practice the number of trees is chosen as a compromise between prediction accuracy and computation time.

2.5.4 Effect of Tree Depth

Tree depth is a measure of how many splits are made before tree growing stops. As mentioned, there are different ways of deciding when to stop growing a tree. One way is to specify the minimum amount of data that should be in each leaf. Another way is to simply specify the tree depth explicitly.

Whereas increasing the forest size does not lead to over-fitting, increasing the tree depth does. In a way, the effect of the tree depth can be compared to that of the number of Gaussians used in GMR. Having a deep tree may cause an over-partitioning of the data. This means that regression will be performed on small clusters which makes it sensitive to variations in the data that may just be due to noise. In Figure 2.7 the effect of different tree depths is visualized. To the left regression using an over-fitting forest is shown and to the right regression by an under-fitting forest is illustrated. Again 200 trees were used with random node optimization and the constraint to the tree depth, that if less than 10 data points reached a node it was declared a leaf node.

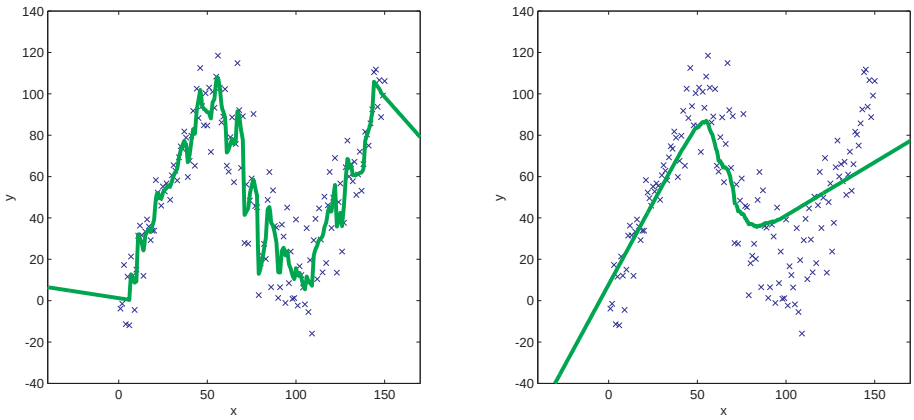


Figure 2.7: Illustration of the impact of tree depth on the regression. The same data as in Figure 2.6 has been used. **Left:** A forest with a maximum tree depth of 6 has been used. **Right:** A forest with a maximum tree depth of 2 has been used.

As with the number of components in a GMR model, the MSPD measure (Equation 2.21) can be used to find the optimal tree depth. When bagging is used to induce tree randomness, the data that is not used at each tree (so-called out-of-bag data), can be used to calculate the MSPD.

2.6 Local Binary Pattern Feature Representation

Because MRI intensity values can vary from scan to scan even for the same tissues, it makes sense to try and find a representation of the MRI images that is invariant to the scaling of the grey-values (grey-scale invariance). A way of achieving this is to cast the MRI images into a feature space where each voxel is characterized by the appearance of its neighbourhood.

The Local Binary Pattern (LBP) is a tool for describing 2D textural information in images in a grey-scale invariant way [33, 34]. It consists of comparing each pixel to the pixels in a circular neighbourhood around it. Let g_c denote the pixel intensity of a center pixel and g_p the pixel intensity of the pth pixel in the circular neighbourhood around the center pixel, then a comparison can be done using the binary decision function, s:

$$s(g_p - g_c) = \begin{cases} 1, & g_p - g_c \geq 0 \\ 0, & g_p - g_c < 0 \end{cases} \tag{2.26}$$

The center pixel can now be described by the binary number resulting from the comparisons with its P neighbouring pixels. Typically interpolation is used to account for the fact that the neighbouring pixels may not lie exactly on the circumference of the circle. The binary number can be converted to a decimal number, which then encapsulates the information about the neighbourhood of the center pixel. This is the LBP of that pixel [34]:

$$LBP(g_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{2.27}$$

where P denotes the number of pixels in the circular neighbourhood. P can be varied as well as the radius, R, of the circle defining which neighbours to use. Using multiple values of P and R, one can achieve a description of each pixel and its neighbourhood at multiple spatial scales. In Figure 2.8, the generation of a LBP is visualized. In the left grid, the center pixel value g_c and its 8 circular neighbours g_0-g_7 are shown. In the middle grid, the value of the center pixel (50) is shown in the centre and the value of its 8 circular neighbours are shown around it (assume an interpolation has been done to find the value of the surrounding pixels). In the right grid, the result of the evaluation with the binary decision function is shown. Because $85 > 50$ a value of 1 is assigned to the top left neighbour pixel. $42 < 50$, so a value of 0 is assigned to the top centre pixel and so on. Finally the binary number is converted to a decimal number using Equation 2.27.

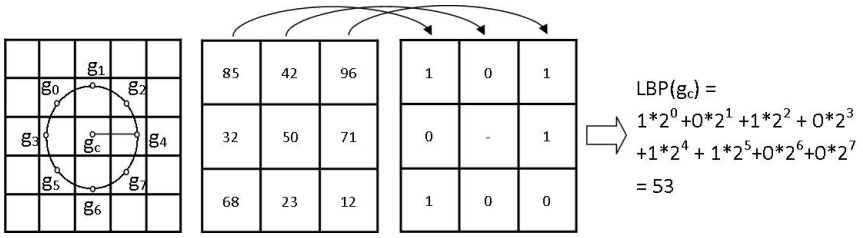


Figure 2.8: Illustration of LBP. The values g_c and g_0-g_7 are the center and neighbourhood pixel values, respectively, shown in the left grid. In the middle grid example values of the center and neighbourhood pixels are shown. Here $g_c = 50$, $g_0 = 85, g_1 = 42, g_2 = 96$ etc. In the right grid, a binary decision function is used to assign values of 1 or 0 to the neighbourhood pixels and the resulting binary number is converted to a decimal number, yielding the LBP.

2.7 Evaluation of Results

There are a number of different ways to quantify the quality of the predicted pCTs. In this report, three methods have been chosen. These will be described in the following sections.

2.7.1 Prediction Deviation

From a model optimization/selection point of view a natural choice is to look at voxel-wise deviations from the desired output. The mean squared prediction deviation has already been introduced as a tool for model optimization. When calculated on the entire image, this measure is well suited for screening of suitable models. To look at an overall estimate of error, however, is not always informative, since deviations are expected to vary throughout different intensity ranges. For instance, it might be more interesting to look at the deviation in the bone intensity range, since this is where MRI is known to be inferior to CT. Also, when adding MRI sequences such as the Dixon, it may be interesting to look at the deviations in specific intensity ranges, to investigate in what situations (if any) the sequence may provide valuable information for the regression models. For these reasons it makes sense to calculate prediction deviations in bins, so as to quantify the error in smaller ranges. This can be achieved by sorting the CT values (measured in Hounsfield unit (HU)) in ascending order, keep track of which pCT voxels correspond to which CT values and then calculate the error in bins of 20 HU.

Due to the square operation in the MSPD it is quite sensitive to outliers. A measure which is less sensitive is the mean absolute prediction deviation (MAPD):

$$MAPD = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{2.28}$$

where N is the data size, y_i is the i'th true output value of the test data and $\hat{y}_i$ is the i'th predicted value when applying the regression model on the test data. The MAPD measures absolute error and thus cannot tell whether the error is attributed to an over- or underestimation of the real value. One may also be interested in looking at whether the model continuously over-/underestimates or if the error is more random in nature. For this the mean prediction deviation (MPD) can be used:

$$MPD = \frac{\sum_{i=1}^{N} y_i - \hat{y}_i}{N} \tag{2.29}$$

2.7.2 Geometric Evaluation

Because correct bone information is crucial for generating DRRs, other than looking at the voxel-wise deviations in the bone intensity range, a geometrical evaluation can be made to measure how well the pCT catches the bone volume compared to the real CT. For this purpose the Dice similarity measure can be used [35]:

$$DICE = \frac{2(A \cap B)}{A + B} \tag{2.30}$$

Where A is the volume of bone in the pCT and B is the true volume of bone in the real CT. A complete overlap of the two volumes will result in a Dice coefficient of 1, whereas no overlap will result in a coefficient of 0. This method requires that the bone volumes in the pCT and real CT are known. A CT bone segmentation can be done using thresholds and the volumes can then be calculated using the known voxel resolution. Along with the Dice coefficient one may report the percentage of falsely classified bone and missed bone in the pCT to give a more thorough picture of the geometrical accuracy. These measures can be defined as follows:

$$Miss\% = \frac{B - (A \cap B)}{B} \cdot 100\% \tag{2.31}$$

$$False\% = \frac{A - (A \cap B)}{A} \cdot 100\% \tag{2.32}$$

2.7.3 Dosimetric Evaluation

The pCT is supposed to be used instead of CT for dose planning in radiation therapy. For that reason, it is sensible to compare a pCT dose plan with a CT dose plan in terms of the estimated delivered dose to organ at risks (ORs) and tumour volumes. A 3D dose plan contains a large amount of dosimetric data and as such it can be hard to compare one against another. The dose-volume histogram (DVH) provides a clinically well established tool for condensing the data and making it easier to interpret and compare [1]. Once a dose plan has been made, an organ can be divided in a dose grid of e.g. 144 $5 \times 5 \times 5$ mm voxels. The differential DVH is then constructed by counting the number of voxels that falls into given dose bins, see Figure 2.9, left. A cumulative variation of the DVH can be constructed by letting each bin represent a percentage of volume that receives a dose equal to or greater than a given dose, illustrated in Figure 2.9, right.

A strategy for making a dosimetric evaluation of the pCT is then be to make a dose plan in a standard way. At Herlev Hospital, this involves delineating

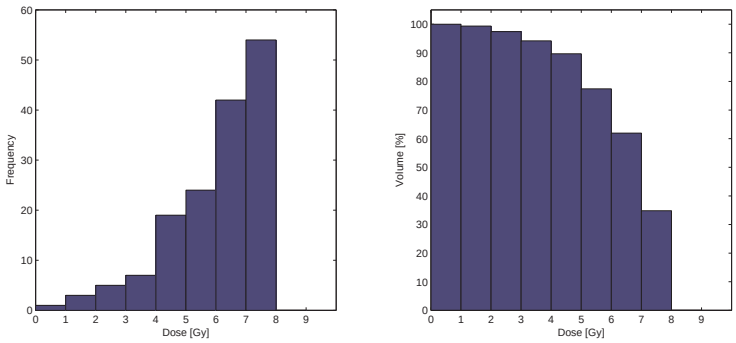


Figure 2.9: Illustrative example of dose-volume histograms. **Left:** Differential DVH. One voxel received between 0 and 1 Gy and 54 voxels received between 7 and 8 Gy. **Right:** Cumulative DVH. 100 % of the volume received 0 Gy or more and 35 % received 7 Gy or more.

structures on a T_1 weighted MRI image set, transferring the structures to the CT, optimize the plan and calculate the absorbed dose. The optimized plan is then transferred to the pCT and the absorbed dose recalculated using the same radiation intensities. The DVHs can then be compared to investigate if the pCT overestimates or underestimates the delivered dose in relevant volumes such as the planning target volume (PTV) and planning risk volumes (PRVs). Besides visually comparing the DVHs, specific points on the DVH can be used to quantify the performance of the pCT. Four DVH points have been chosen in this report. These are $D_{98\%}$, D_{median}, $D_{2\%}$ and D_{mean}; the absorbed dose received by 98 %, 50 %, 2 % of the volume and the mean absorbed dose, respectively. $D_{98\%}$, D_{median} and $D_{2\%}$ are values relevant to the planning target volume (PTV) and $D_{2\%}$ and D_{mean} are values relevant to the planning risk volume (PRV). These values were chosen in accordance with recommendations by The International Commission on Radiation Units & Measurements (ICRU) [36].

Chapter 3

Methods and Procedures

3.1 Data Acquisition and Preprocessing

During the project, 3 patients receiving palliative cranial radiation therapy were scanned with the dUTE and multi-echo Dixon (mDixon) sequence along with a standard T_1 weighted sequence and CT. Prior to participating, written consent was obtained from the patients. After acquiring the data, all scans were anonymized using Conquest Dicom Server software. From here on the patients were named *h01*, *h02* and *h03* for easy distinction.

3.1.1 Scanner Set-up

Appropriate patient fixation devices were used for all scans. All CT images were acquired on a Philips Brilliance Big Bore CT using standard settings. All MRI images were acquired on a Philips Panorama 1 Tesla open MRI-scanner using a Flex-coil. Two dUTE acquisitions were made. One with a flip angle of $\alpha = 10°$ and one with a flip angle of $\alpha = 25°$. The mDixon sequence recorded in-phase, out-of-phase, water-only and fat-only image sets. The water-only and fat-only image sets were used in this project. The parameters for both the dUTE and

mDixon sequence are summarized in Table 3.1 along with image information about all acquired images.

	T_{E_1} [ms]	T_{E_2} [ms]	T_R [ms]	dT_E [ms]	Voxel resolution $[mm^3]$	FOV [mm]
dUTE	0.09	3.5	7.1	-	$1 \times 1 \times 1$	256
mDixon	6.9	-	16	3.5	$1 \times 1 \times 1.5$	250.5
CT	-	-	-	-	$0.6 \times 0.6 \times 2$	220

Table 3.1: Acquisition parameters and image information. T_{E_1} is the first echo time of the dUTE and mDixon sequences. T_{E_2} is the second echo time of the dUTE sequence. T_R is the repetition time. dT_E is the time interval between echo acquisitions in the multi-echo Dixon sequence. FOV is the field of view.

3.1.2 Registration

All MRI images were acquired at the same set-up and were thus internally registered. The CT images were rigidly co-registered with the MRI images using the in-built mutual information algorithm of 3D Slicer (www.slicer.org). An initial geometrical alignment was done prior to applying the algorithm. 50 bins were used to construct the joint histogram and linear interpolation was applied to the transformed image. The result was visually inspected to ensure that the registration had worked appropriately. All images were then re-sliced to the resolution of the dUTE images and cropped to the smallest field of view (FOV).

Due to a large artefact in the MRI image sets of patient h02 (probably susceptibility artefact due to dental fillings), most of the bottom part of the head (jaw region) was cropped on all images for this patient. This was done as to not corrupt the training data of the models.

3.1.3 Local Binary Pattern-like Feature Extraction

Inspired by the approach by Pauly *et al.*, who used an LBP-like feature for predicting organ bounding boxes in Dixon MRI image sets [17], a similar 3D version of LBPs was implemented for this project. The 3D version of LBP differed from the 2D version in a number of ways. First of all the the neighbourhood voxels was not defined by a circular (2D) neighbourhood but by a spherical (3D) one. This meant that 26 neighbouring voxels were used to create each LBP value.

Also, instead of doing a voxel by voxel neighbourhood comparison, mean values of cuboidal regions were compared. This meant that for each center voxel, the mean value of a cube centred on that voxel was computed and compared to the mean values of cubes centred on the spherical neighbourhood voxels. In Figure 3.1, this is visualized for a 2D example. For fast calculation of the cube mean values, integral volume processing was implemented [37]. The 26 digit binary number, resulting from the mean value comparisons, was then converted to a decimal number in the same way as in the 2D method.

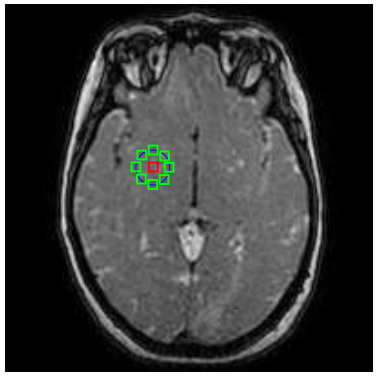


Figure 3.1: Example of the modified LBP implementation. To calculate the LBP in a voxel, the mean value of the voxel intensities in the red box centred on that voxel is compared to mean values of the voxel intensities in the green boxes centred on neighbouring voxels. Note that this is a 2D simplification. In the actual implementation the boxes are 3D cubes and the neighbourhood is defined by a sphere, not a circle.

The 3D LBP was implemented to use 3 different spatial scales. This was achieved by using a cube size of $3 \times 3 \times 3$ mm and a sphere radius of 3 mm for the first scale and then adding 2 and 4 mm to those values for the second and third scale, respectively. The result of LBP feature extraction from one image set was thus 3 new image sets each consisting of the LBPs at a given scale. In Figure 3.2, the LBP feature images corresponding to the image in Figure 3.1 can be seen.

3.1.4 Filtering

Following the approach of Johansson *et al.* [5], filtered counterparts of all MRI images were created. Each image was mean filtered by assigning each voxel a value corresponding to the mean value of the voxels in a 26 voxel cuboidal neighbourhood around that voxel. Standard deviation filtering was achieved in

Figure 3.2: LBP feature images. The feature images correspond to the image shown in Figure 3.1. In the left panel, a sphere with radius $3mm$ and cubes of size 3^3 mm^3 were used. In the middle panel the sphere radius was $5mm$ and cube size 5^3 mm^3. In the right panel the sphere radius was $7mm$ and cube size 7^3 mm^3.

a similar fashion, but this time finding the standard deviation of the 26 voxel neighbourhood. The filtering thus resulted in two new image sets per original MRI image set, see Figure 3.3. All filtering was done in Matlab using scripts developed by the author. Whenever an image set was used in a model so was its filtered counterpart. The only exception to this, was the LBP feature image sets.

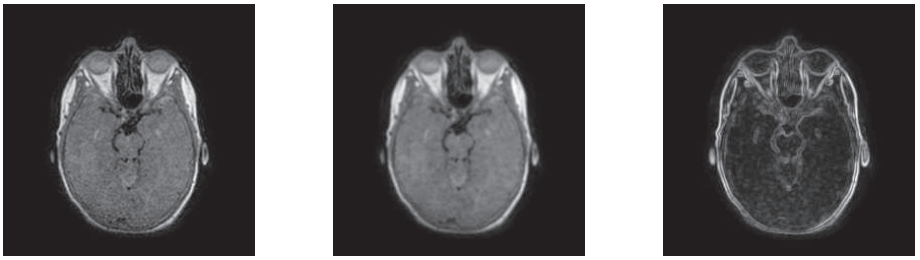


Figure 3.3: Example of the filtered images. **Left:** A UTE image prior to filtering. **Middle:** Mean filtered image. **Right:** Standard deviation filtered image.

3.1.5 Air Mask

To reduce the amount of data and the computational burden when training the models a mask was created which removed the air surrounding the head in all images. This was achieved using Matlab built-in functions. A global threshold was found to binarize the image in to two classes; a low intensity region (air) and a high intensity region (foreground). Using this method some foreground

areas were classified as air, so morphological operations were applied to remove these spots. Lastly, connected component analysis was used to determine which region was foreground and which was air. The air mask was derived from the mean filtered version of echo 1 of the high angle dUTE image and applied to all corresponding image sets. The result was visually inspected to ensure that no important structures were removed. An example of the air mask is shown in Figure 3.4.

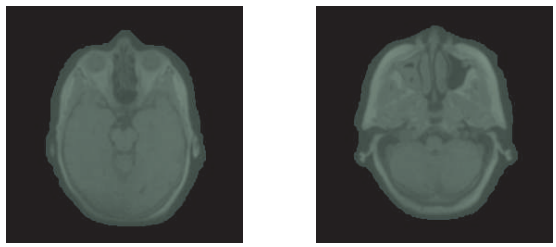


Figure 3.4: Example of the air mask. The green area is the air mask and everything outside this area is excluded from the images.

3.1.6 Defining the Input/Output Data for the Models

Once the pre-processing was done, the training data for the models was defined. Models with dUTE only, with dUTE+mDixon and with dUTE+LBP input were to be trained. The dUTE only model consisted of 12 input variables, namely a dUTE acquisition at 10° (2 image sets), a dUTE acquisition at 25° (2 image sets) and their filtered images (8 image sets). The dUTE+mDixon model had in addition to the aforementioned images, an mDixon water and an mDixon fat image set and their filtered versions as input thus yielding 18 input variables in total. The model with dUTE+LBP consisted of the same input variables as the dUTE only model but with the addition of 3 LBP image sets per original dUTE image set. This resulted in a total of 24 input variables.

The output variable was in all cases the CT image set.

3.2 Gaussian Mixture Regression

A Matlab implementation of Gaussian mixture regression, [38], was used throughout this project. The implementation utilized the built-in k-means function of the Matlab Statistics toolbox to give an initial estimate of the component mean

values and covariance matrices for the EM algorithm. Initial seeds for the k-means algorithm were chosen at random. EM was stopped when the increase in log likelihood was smaller than 10^{-10}.

3.2.1 mDixon Input Study

To evaluate the impact of using mDixon images as an additional input for the GMR model, dUTE only and dUTE+mDixon models were trained for comparison. A leave-one-out approach was taken. This meant that each model was trained on data from two patients and then applied to the third to predict a pCT. In this way three pCTs were predicted from dUTE+mDixon models and three from dUTE only models. Following the method of Johansson *et al.*, [5,16], 20 components were used for estimating all GMMs.

3.3 Random Forest Regression

The Matlab built-in implementation called `TreeBagger`, was used for RaFR. With the parameter settings used in this project, `TreeBagger` is identical to Breiman's version of the Random Forest framework as described earlier. For randomization of the individual trees, a combination of bootstrap aggregating (*bagging*) and random node optimization was used. This meant that only a fraction of the total training data was used to train each tree and that only one third of the input variables were chosen at random at each node to determine the best split.

3.3.1 Model Optimization Study

In this study the optimal tree depth for pCT generation was investigated. In the Matlab RaF implementation, the tree depth was controlled implicitly by the minimum leaf size parameter. This parameter controlled the minimum amount of data that should be in a node before it was declared a leaf. As such, the user had no direct control over the tree depth. The Matlab implementation provided a method for estimating the optimal minimum leaf size. This consisted of storing which data was out-of-bag (i.e. the data not used to train each tree) for each tree and then using this as test data to make an estimate of the prediction error. Using this method, 5 regression forests each with 25 trees were trained on one patient (h01) and the out-of-bag data was used to calculate the MSPD

as a function of the number of trees. The input variables were dUTE+mDixon image sets. The minimum leaf size was set to 5, 10, 20, 50 or 100 in the forests for comparison.

3.3.2 LBP Input Study

In this study the effect of using the 3D LBP feature image sets as an input to the RaFR model was investigated. Three RaFs were trained on data from one patient (h01) using 25 trees in each forest and a leaf size of 10. One forest had the dUTE image sets as input, one had the dUTE image sets along with their corresponding LBP feature image sets as input and one had the dUTE+mDixon image sets as input. The models were tested on data from another patient (h03) to predict the pCTs.

3.3.3 mDixon Input Study

As in the mDixon input study of the GMR model the impact of using mDixon images in the RaFR model was evaluated by training dUTE only and dUTE+mDixon models for comparison. Again a leave-one-out approach was used to estimate three pCTs using mDixon images and three without them. A forest consisting of 80 trees with a minimum a leaf size of 10 was used.

3.4 Post-processing

3.4.1 Prediction Deviation

For the mDixon input studies and LBP input study, the MAPD and MPD in bins of 20 HU were calculated to quantify the prediction errors made in specific intensity ranges of the real CT image set when using the additional input and when not using the additional input. Furthermore the MAPD for the entire image sets and for the voxels inside the air mask was calculated in order to be able to compare the overall accuracy of the different models. These computations were all carried out in Matlab.

3.4.2 Geometric Evaluation

The results from the mDixon input studies were evaluated geometrically using the Dice coefficient and the $Miss\%$ and $False\%$. These values were computed by importing the pCTs and real CTs into Varian Eclipse Treatment Planning System software. Here, the bone voxels in the image sets were automatically segmented. The software provided tools for finding the intersection between the bone regions of the pCT and the real CT, as well as for measuring the volume of these regions. These quantities were used to calculate the Dice coefficient and $Miss\%$ and $False\%$ as described earlier.

3.4.3 Dose Plan and DVHs

For the mDixon input studies, a dosimetric evaluation of the pCTs of patient h03 was performed. The cerebrum, eyes and Hippocampus were automatically segmented on the T_1 weighted image set. A Volumetric Modulated Arc Therapy (VMAT) dose plan was set up and optimized in Varian Eclipse Treatment Planning System following the standard clinical procedure at Herlev. The cerebrum was the planning target volume (PTV) and the eyes and Hippocampus were planning risk volumes (PRVs). 30 Gy was to be delivered in 10 fractions using 2 arcs with an output of 286 and 285 Monitor Units (MU), respectively. The simulated delivered dose was calculated using the CT (considered the true delivered dose) and pCTs and DVHs of the PTV and PRVs were made. $D_{98\%}$, D_{median}, $D_{2\%}$ and D_{mean} DVH points were noted. The dose plan was pre-clinical, in the sense that no oncologist verified the dose distributions or the appearance of the automatically segmented structures.

Chapter 4

Results

The results are presented in 3 sections. One with the result of the RaF model optimization, one which focuses on the impact of using the LBP feature in RaFR and one which focuses on the impact of the mDixon input for both GMR and RaFR.

4.1 Random Forest Model Optimization

In Figure 4.1 the MSPD has been plotted as a function of the number of trees in a forest. 5 forests were trained with different terminal leaf size. This study was carried out to decide which leaf size should be used in the later studies. No marked difference was observed at 25 trees, when using a leaf size of 5 or 10, so 10 was chosen because shallower trees should be faster to train than deeper ones. It is seen that as the number of trees increases, the MSPD decreases and converges towards a constant value.

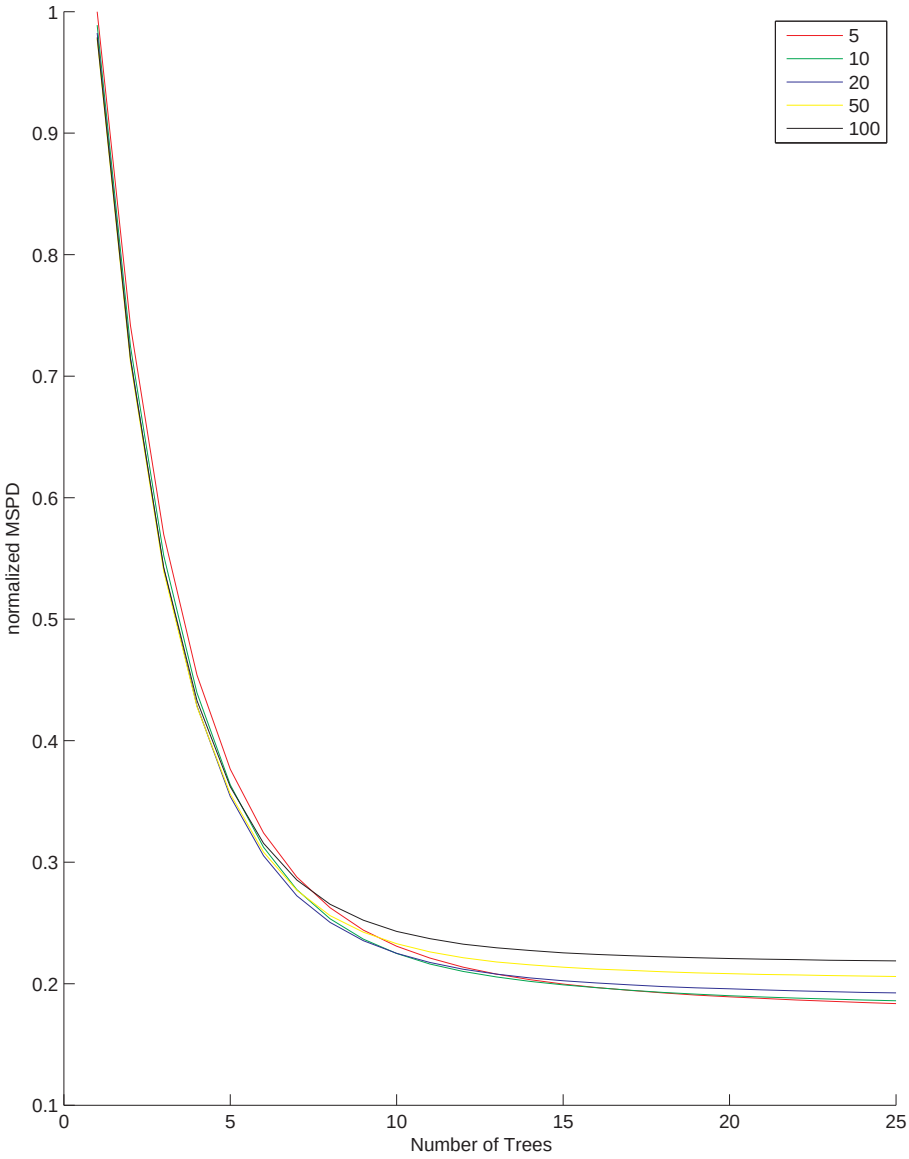


Figure 4.1: The mean squared prediction deviation as function of the number of trees in the forest for five forests with leaf sizes of 5, 10, 20, 50 or 100.

4.2 Random Forest LBP Input

In Figure 4.2 the MAPD and MPD have been plotted as a function of the real CT value (calculated in bins of 20 HU) for comparison of a model with dUTE only input and one with dUTE+LBP input. It is observed, that in the air region (-1000 HU), the MAPD is lowest for dUTE+mDixon and highest for dUTE+LBP. In the range −700-−100 HU the MAPD is lowest for dUTE+LBP and highest for dUTE+mDixon. In the range 500-1500 HU the MAPD is lowest for dUTE+mDixon and highest for dUTE only. It is observed from the MPD, that the differences in MAPD are mainly due to systematic errors, i.e. a consistent overestimation of air and underestimation of bone (in the 700+ HU range). In table 4.1, the MAPDs are shown for the three pCTs. It is observed that the error is lowest for dUTE+mDixon and highest for dUTE+LBP when looking at the entire image sets. When only looking in the region inside the air mask, the error is still lowest for dUTE+mDixon but highest for dUTE only. It is observed that including the air region increases the MAPD for dUTE+LBP and decreases it for dUTE+mDixon and dUTE only.

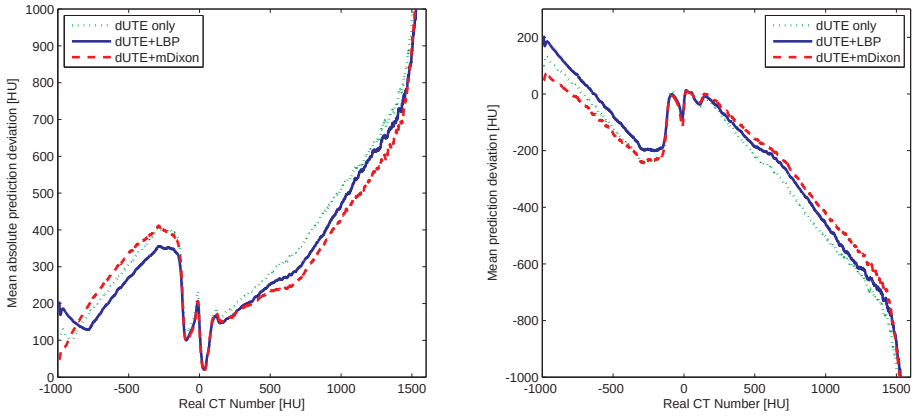


Figure 4.2: The MAPD (left) and MPD (right) as a function of the real CT value (calculated in bins of 20 HU) for models using dUTE only, dUTE+LBP and dUTE+mDixon input.

	dUTE only	dUTE+LBP	dUTE+mDixon
MAPD, entire image	121.7 HU	184.3 HU	81.1 HU
MAPD, inside air mask	145.9 HU	133.0 HU	122.6 HU

Table 4.1: The mean absolute prediction deviation for the entire image sets and inside the air mask for models using dUTE only, dUTE+LBP and dUTE+mDixon input.

4.3 mDixon Input Study

The results of the mDixon input variable studies for both models have been collected in this section.

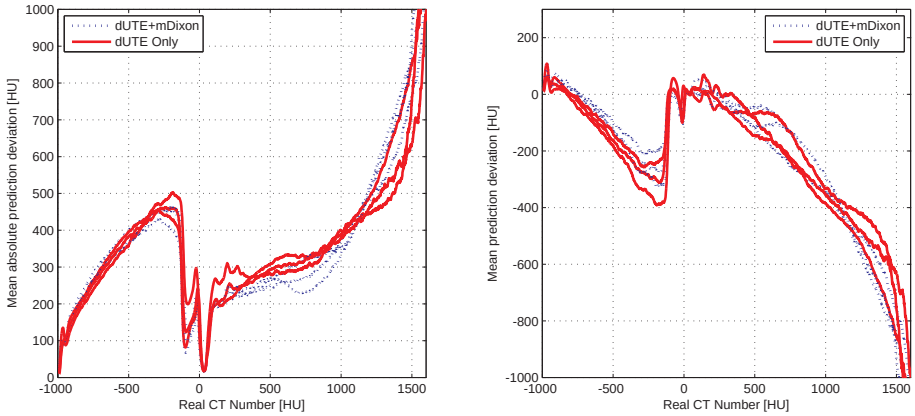
4.3.1 Prediction Deviations

In Figure 4.3 the MAPD and MPD have been plotted as a function of the real CT value (calculated in bins of 20 HU) for the three patients. In the top panels (4.3a) the results for GMR with and without mDixon are shown and in the bottom panels (4.3b) the same are shown for RaFR.

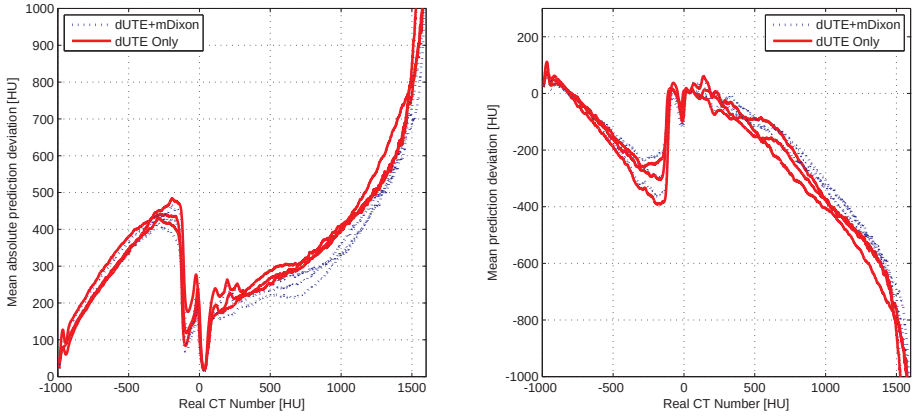
Looking at the top panels (GMR), similar patterns are observed both with and without the mDixon: the absolute error is observed to be low when predicting CT values around −1000 HU and 0 HU. Looking at the impact of using dUTE+mDixon, for two patients the MAPD is observed to be lower in the 500-1000 HU range for the dUTE+mDixon pCTs compared to the dUTE only. In the 1000-1500 range the opposite is the case. The MPD, which as mentioned can quantify the systematic errors, reveals that the absolute error seen in the ranges from −1000 HU to −100 HU and 250-1500 HU is mainly due to a consistent underestimation of the real CT value.

Looking at the bottom panels (RaFR), a lower MAPD is observed for the dUTE+mDixon compared to the UTE only in the 500-1500 HU range. Looking at the MPD, it is observed that the systematic error is also lower (closer to 0) for the UTE+mDixon compared to the UTE only in the 500-1500 HU range.

In Table 4.2, the mean absolute prediction deviation is shown for all predicted pCTs using RaFR and GMR with dUTE+mDixon or dUTE only as inputs. Results calculated for the entire image sets (top half) and inside the region of the air mask (bottom half) are shown. When looking at the entire image set, on average GMR with dUTE only has the lowest overall absolute error followed by RaFR with dUTE+mDixon. It is seen that RaFR with dUTE+mDixon has the lowest standard deviation. Looking at the the region inside the air mask, on average, RaFR with dUTE+mDixon has the lowest error, followed by GMR with dUTE+mDixon, RaFR with dUTE only and GMR with dUTE only.



(a) GMR



(b) RaFR

Figure 4.3: The MAPD (left) and MPD (right) as a function of the real CT value (calculated in bins of 20 HU) for the pCTs for the three patients. The top panels are for GMR and the bottom panels are for RaFR.

Entire image set				
	GMR dUTE+ mDixon	**GMR** dUTE only	**RaFR** dUTE+ mDixon	**RaFR** dUTE only
MAPD, h01	68.7 HU	67.9 HU	70.0 HU	69.5 HU
MAPD, h02	75.0 HU	73.4 HU	70.0 HU	88.2 HU
MAPD, h03	70.3 HU	65.6 HU	69.3 HU	78.7 HU
Mean value	71.3(±2.65)HU	69.0(±3.25)HU	69.8(±0.32)HU	78.8(±7.62)HU
Inside air mask				
	GMR dUTE+ mDixon	**GMR** dUTE only	**RaFR** dUTE+ mDixon	**RaFR** dUTE only
MAPD, h01	127.3 HU	133.4 HU	112.2 HU	125.6 HU
MAPD, h02	128.1 HU	142.1 HU	112.8 HU	137.6 HU
MAPD, h03	128.1 HU	139.3 HU	113.2 HU	134.5 HU
Mean value	127.8(±0.4)HU	138.3(±3.6)HU	112.7(±0.4)HU	132.6(±5.1)HU

Table 4.2: The mean absolute prediction deviation (MAPD) calculated for each predicted pCT using the different regression models. Also the mean MAPD for each method with one standard deviation confidence interval is shown. Values calculated for the entire image sets and for voxels inside the air mask are shown.

4.3.2 Geometric Evaluation

In Table 4.3 the $Miss\%$, $False\%$ and Dice score for the individual patients along with the average when using GMR and RaFR is shown. It is observed that the general tendency for the individual patients is a lower $Miss\%$ and $False\%$ and a higher Dice score when using dUTE+mDixon compared to dUTE only. This is the case for both GMR and RaFR and is further confirmed by the mean values computed over all three patients. It is also observed that on average, the pCTs predicted using RaFR with dUTE+mDixon input have the lowest $Miss\%$ and $False\%$ and highest Dice score. This is also the method that shows the lowest standard deviation. On average GMR using dUTE only produces the highest $Miss\%$ and $False\%$ and lowest Dice score.

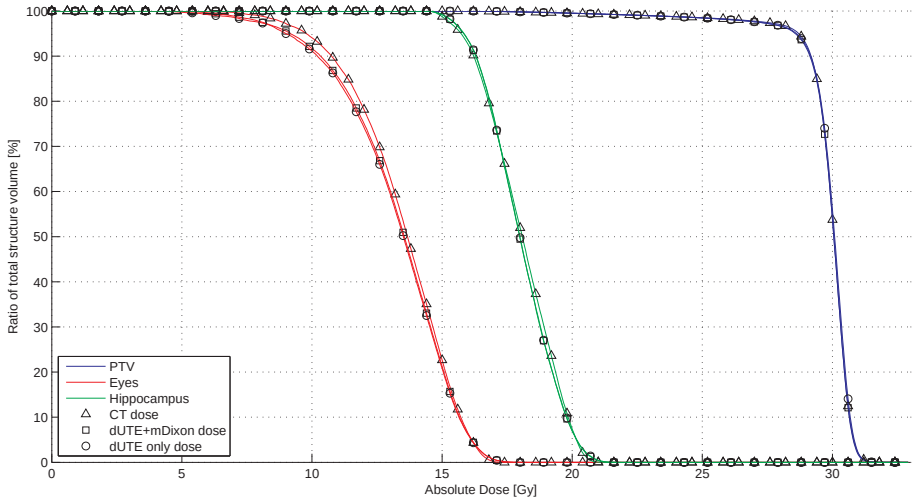
GMR		
	dUTE + mDixon	dUTE only
h01, $Miss\%$	18.3%	26.0%
h02, $Miss\%$	16.7%	17.4%
h03, $Miss\%$	23.4%	27.3%
h01, $False\%$	31.5%	31.1%
h02, $False\%$	26.2%	33.4%
h03, $False\%$	29.0%	30.8%
h01, Dice	0.75	0.71
h02, Dice	0.78	0.74
h03, Dice	0.74	0.71
Mean $Miss\%$	19.5%(±3.48%)	23.6%(±5.44%)
Mean $False\%$	28.9%(±2.63%)	31.7%(±1.42%)
Mean Dice	0.76(±0.38)	0.72(±0.02)
RaFR		
	dUTE + mDixon	dUTE only
h01, $Miss\%$	20.3%	25.8%
h02, $Miss\%$	13.9%	16.1%
h03, $Miss\%$	17.5%	24.9%
h01, $False\%$	22.1%	26.8%
h02, $False\%$	22.4%	30.2%
h03, $False\%$	22.3%	28.7%
h01, Dice	0.79	0.74
h02, Dice	0.82	0.76
h03, Dice	0.80	0.74
Mean $Miss\%$	17.2%(±3.22%)	22.0%(±5.14%)
Mean $False\%$	22.3%(±0.15%)	28.6%(±1.75%)
Mean Dice	0.80(±0.01)	0.75(±0.01)

Table 4.3: Bone miss percentage, false bone percentage and Dice score for GMR and RaFR with and without mDixon input for the three patients. The values below the dashed lines are the mean values calculated over all three patients with one standard deviation confidence limits in parentheses.

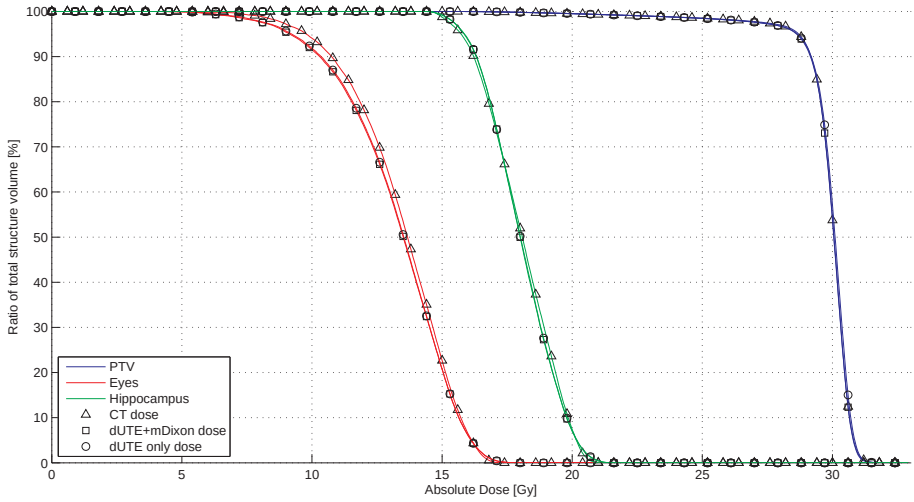
4.3.3 Dosimetric Evaluation

The dosimetric evaluation was performed for patient h03. In Figure 4.4, the dose-volume histograms of the planning target volume (PTV) and planning risk volumes (PRVs) (eyes and Hippocampus) for CT and pCT based dose calculations are shown. From a visual inspection, it is observed that the largest differences are present in the eye volume for both RaFR and GMR. It is hard to see if there are any differences between using dUTE+mDixon or dUTE only.

In Table 4.4, the DVH points, $D_{98\%}$, D_{median}, $D_{2\%}$ and D_{mean}, are shown for the three volumes of Figure 4.4 along with the percent deviation between the points estimated from the pCT based doses and the CT based doses. It is observed that all deviations are smaller than 2 %. When a deviation is observed in the PTV it is due to an underestimation of the true dose at $D_{98\%}$ and due to an overestimation at D_{median} and $D_{2\%}$. For the eyes and Hippocampus, the deviation in D_{mean} is due to an underestimation of the true dose and in $D_{2\%}$ it is due to an overestimation.



(a) Comparison of GMR models.



(b) Comparison of RaFR models.

Figure 4.4: Dose-volume histograms. The cerebrum was the planning target volume (blue) and the eyes (red) and hippocampus (green) were the planning risk volumes. The CT calculated dose is marked with triangles for comparison with the dose calculated based on pCTs from models using dUTE+mDixon (squares) and dUTE only (circles).

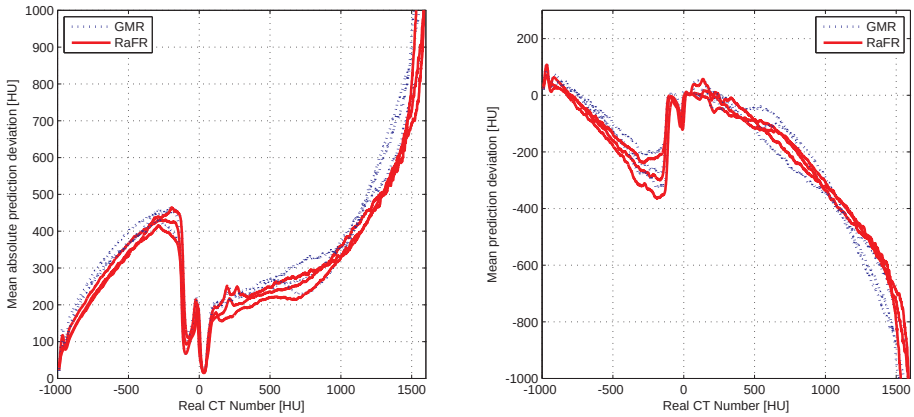
	CT	**RaFR** dUTE+ mDixon	**RaFR** dUTE only	**GMR** dUTE+ mDixon	**GMR** dUTE only
PTV, $D_{98\%}$	26.49 Gy	26.28 Gy	26.31 Gy	26.25 Gy	26.28 Gy
% deviation	-	0.79 %	0.68%	0.91 %	0.79%
PTV, D_{median}	30.05 Gy	30.06 Gy	30.10 Gy	30.05 Gy	30.09 Gy
% deviation	-	-0.03 %	-0.17 %	0.00 %	-0.11 %
PTV, $D_{2\%}$	30.99 Gy	30.99 Gy	31.02 Gy	30.99 Gy	31.02 Gy
% deviation	-	0.00 %	-0.10 %	0.00 %	-0.10 %
Eyes, D_{mean}	13.42 Gy	13.20 Gy	13.22 Gy	13.23 Gy	13.17 Gy
% deviation	-	1.65 %	1.45%	1.39%	1.83%
Eyes, $D_{2\%}$	16.50 Gy	16.59 Gy	16.59 Gy	16.59 Gy	16.59 Gy
% deviation	-	-0.55 %	-0.55%	-0.55%	-0,55%
Hippocampus, D_{mean}	18.04 Gy	18.02 Gy	18.03 Gy	18.00 Gy	18.01 Gy
% deviation	-	0.12%	0.06%	0.19%	0.16%
Hippocampus, $D_{2\%}$	20.43 Gy	20.58 Gy	20.58 Gy	20.55 Gy	20.58 Gy
% deviation	-	-0.73%	-0.73%	-0.59%	-0.73%

Table 4.4: Comparison of DVH points for the planning target volume (PTV), eyes and Hippocampus using the different models. The percent deviations from the dose calculated using the real CT are also shown.

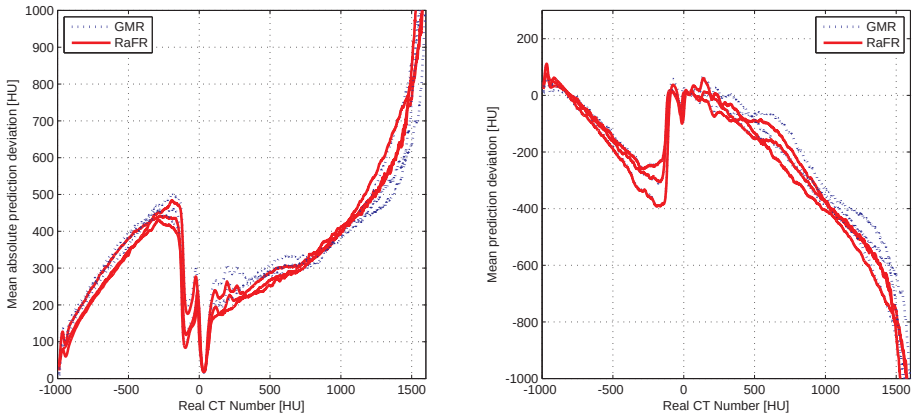
4.3.4 Comparing GMR and RaF

In Figure 4.5a the MAPD and MPD have been plotted as a function of the real CT value (calculated in bins of 20 HU) for RaFR and GMR both using dUTE+mDixon input. It is observed that the MAPD is lower for RaFR for high density CT values (1000-1500 HU). Also the MPD is closer to 0 in that range for RaFR. It is seen that in the range from approximately -700 HU to -400 HU the MPD is closer to zero for GMR. In Figure 4.5b the same is shown as in Figure 4.5a, but for models with dUTE only. Here, it is observed that the MAPD is lower for two of the patients for CT values of 1000+ HU using GMR.

In Figure 4.6, the pCTs and real CT from patient h03 are shown for three axial slices through the head. It is observed that noise is present in the air region of the pCTs predicted using dUTE+mDixon. A few differences between the pCTs and the real CT should be noted. First of all, the fixation device seen in the real CT is not present in the pCTs. Secondly, the streak artefact due to dental fillings seen in the real CT is not present in the pCTs (although the tooth from which it originates has a different intensity than the rest of the teeth). In the middle row of Figure 4.6, it is seen that to various degrees, the pCTs suffer from errors in the nasal region, the paranasal sinuses and in the ear region.



(a) With mDixon input.



(b) Without mDixon input.

Figure 4.5: Comparison of RaFR and GMR, with and without mDixon input. The MAPD (left) and MPD (right) plotted as a function of the real CT value (calculated in bins of 20 HU).

(a) Real CT **(b)** pCT, RaFR dUTE+mDixon **(c)** pCT, RaFR dUTE only **(d)** pCT, GMR dUTE+mDixon **(e)** pCT, GMR dUTE only

Figure 4.6: Columnwise visual comparison of the pCTs and the real CT from patient h03. The top row shows an inferior (lower) axial slice of the head. The middle row shows a medial axial slice. The bottom row shows a superior (upper) axial slice.

Chapter 5

Discussion

5.1 Random Forest Model Optimization Study

The purpose of this study was to find a suitable tree depth for pCT prediction using RaFR. As mentioned, this involved looking at the minimum leaf size. The mean squared prediction deviation (MSPD) as a function of the number of trees was shown in Figure 4.1 for forests with a minimum leaf size of 5, 10, 20, 50 or 100 data points. It was observed that using 5 or 10 data points produced the lowest MSPD using 25 trees, so 10 was chosen as the optimal minimum leaf size due to the theoretically faster training and testing using shallow trees compared to deep ones. Figure 4.1 also serves to demonstrate one of the properties of the RaF framework: as the number of trees increases the prediction error decreases towards a constant value. The out-of-bag data for each tree was sampled with replacement so as the number of trees increased, the number of weak predictions made with the same input data also increased. Therefore, what is shown is that as the number of weak predictions increases the final prediction error decreases. The optimization study was meant as a preliminary study to find an estimate of the optimal tree depth before training the final models. The training was performed on data from one patient and with 25 trees in each forest. This may pose a problem to the generalisability of the results, i.e. the data from one patient may not be representative of the entire population of data and the trend in the MSPD curves may change for forests of more than 25 trees. This, of

course, should be considered before drawing any final conclusions on the optimal minimum leaf size. A more thorough study, using data from more patients and a forest size closer to the one used for actual pCT prediction could be carried out in the future.

5.2 Random Forest Local Binary Pattern Input Study

This study was carried out in order to investigate if the LBP feature image sets could improve the prediction accuracy of RaFR. It was observed in Figure 4.2 that in the air region (-1000 HU), the model with LBP feature image sets had a larger MAPD and MPD compared to models with dUTE+mDixon and dUTE only input. This can probably be explained by the air mask used to limit the training data amount prior to training the models. Because the LBP of a voxel contains information about the neighbourhood of that voxel, air voxels close to the head have a different LBP value than those far from the head that are surrounded only by other air voxels. When the air mask was applied to the training data, it essentially removed all the air voxels that were spatially far from the head. Because the training data lacked these voxels, the resulting model was not trained to handle their LBP value, resulting in prediction errors in those voxels. The error is clearly seen in Figure 5.1, where a dark "halo" of correctly predicted air values surround the head but further away, the air voxels are overestimated. A solution to this problem could be to assign a bulk intensity value of -1000 HU to all pCT voxels outside the air mask.
In Figure 4.2, it was seen that in the range from approximately -700 HU to -100 HU, the model with dUTE+LBP input had the lowest MAPD. In theory, lung tissue is the only tissue type in this Hounsfield range, so voxels of this intensity should not be present in the head region of the body. From inspection of the real CTs used in this project, it was found that voxels of these intensities were placed in transition regions going from air to tissue. Instead of a sharp transition, a smooth gradient was observed, which lead to intensity values between -1000 HU and -100 HU. These regions of smooth transitions were observed along all such edges, meaning on the entire periphery of the head and in the air cavities inside the head. It was investigated whether the smoothing artefact was due to the interpolation done when registering and reslicing the CT image sets to the dUTE image sets. This was not the case, since it was also present in the unprocessed CT images. It was thus assumed that the artefact was introduced in the CT acquisition phase. From the error curves in Figure 4.2, it seems that the model with the LBP feature is better able to mimic this property of the real CT compared to the models without it.

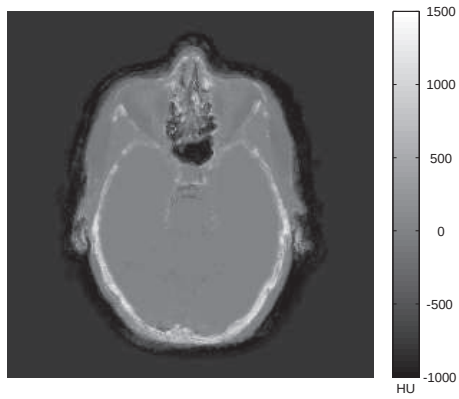


Figure 5.1: Example of air mask related prediction error when using LBP feature image sets. The shown pCT was predicted using dUTE+LBP input. As can be seen, the value of the air voxels far from the head is overestimated.

The three models showed similar error-wise behaviour in the soft tissue region (-100 HU to 500 HU), with the exception that the model with dUTE only input had larger errors in the 500 HU area.

When looking at the error in the bone range (500+ HU), the model with dUTE+LBP input had a performance between that of the models with dUTE only input (worst) and dUTE+mDixon input (best). It was closest to the dUTE+mDixon curve in the range from 500 HU to 1000 HU and the dUTE only curve from 1000 HU to 1500 HU. From this, it seems that adding the LBP feature input improves the prediction accuracy in the bone region compared to using dUTE only.

Due to the prediction errors made outside the air mask when using the dUTE+LBP input, the MAPDs shown in Table 4.1 were calculated both for the entire pCT and for the voxels inside the air mask. When looking at the entire image sets, the relatively large error observed for the model with dUTE+LBP input is mainly caused by all the air voxels far from the head which were overestimated. When looking at the head region only (inside the air mask), it was seen that the model with dUTE+LBP input performs better than the one with dUTE only input and worse than the model with dUTE+mDixon input. This confirms what was observed in the error curves of Figure 4.2.

Summary: It seems to be possible to get better predictions when using the LBP feature image sets as an additional input compared to using dUTE only

input. This may be attributed to the neighbourhood information that these image sets contain. Still, using the dUTE+mDixon input resulted in better overall predictions than using the dUTE+LBP input. An advantage of the LBP feature image sets is that they are derived directly from the dUTE images, so no additional scans are needed. A reduction in scan time is thus obtained compared to using the mDixon input. Given this fact and the results shown, the method looks promising although further studies using more data and a larger forest should be carried out.

5.3 mDixon Input Study

The RaFR LBP input study already suggested that using the mDixon input produced better predictions compared to using dUTE only input. In this study a more thorough investigation of the mDixon input was carried out using more data and a larger forest. Furthermore, the use of GMR for pCT prediction using both dUTE+mDixon and dUTE only input was investigated. All resulting pCTs were evaluated thoroughly using both the prediction deviation, Dice score and dosimetric deviation.

5.3.1 Gaussian Mixture Regression

The piecewise deviations of the pCTs from the real CT value was shown in Figure 4.3a for GMR using dUTE+mDixon and dUTE only input. From the MAPD and MPD error curves it was hard to determine whether the GMR model was improved when using the mDixon input. In one region (500-1000 HU) the predictions were better (for two patients) but in another they were worse (1000-1500 HU).

When looking at the MAPDs in Table 4.2, the difference between using mDixon or not for the entire image sets seems relatively small. When looking in the confined region of the air mask, however, the difference is larger and consistently in favour of using the mDixon input. From visual inspection of the pCTs, it was noticed that when using the mDixon input, predictions in the surrounding air region was more affected by errors than when using dUTE only input (see Figure 4.6). It was also observed that the mDixon image sets themselves were, in general, more influenced by noise in air voxels than the dUTE image sets (the noise in the air voxels can be seen in Figure 2.2). This may be attributed to the post-processing done to these image sets in order to derive the water-only and fat-only image sets. The noise may be the reason that the models with

mDixon input do not show smaller errors than models with dUTE only input when looking at the entire image sets. Again, one could consider assigning a bulk value of -1000 HU to all pCT voxels outside the air mask.

As mentioned in the introduction, the GMR model has previously been used by Johansson *et al.* to predict pCTs [5, 16]. One of the goals of this project was to investigate the robustness of the model and the ability to reproduce the results in a different clinical set-up (e.g. using a 1 T MRI scanner instead of a 1.5 T MRI scanner). Johansson *et al.* reported an average MAPD inside the air mask of 137 HU with a minimum of 117 HU and maximum of 176 HU for the individual patients. Their study included 5 patients and used dUTE input as well as an additional T_2 weighted MRI image set (which was later found to be redundant). In the present study, the average MAPD inside the air mask was found to be 138.3 HU for dUTE only input, which is close to what Johansson *et al.* reported. Johansson *et al.* also did a piecewise evaluation of the error using the MAPD and MPD calculated in bins of 20 HU. However, they calculated the error curves as a function of the predicted CT value and not the real CT value (as done in this report) so a direct comparison was not possible. Still, the overall MAPD results were comparable and also from a visual inspection the results obtained in this project look similar to those reported by Johansson *et al.* The difference in clinical set-up thus seems to not affect the accuracy of the predicted pCTs as long as the GMR model is trained on data from the specific clinical set-up that it should be used in.

The geometrical evaluation showed a tendency in favour of using the mDixon input. Referring to Table 4.3, the bone $Miss\%$ for each patient pCT was consistently lower for models using the mDixon input compared to models with dUTE only input. Except for the pCT of patient h01, the same was the case for the bone $False\%$. Furthermore, the Dice coefficient was higher for all pCTs when using mDixon input, suggesting a better geometrical accuracy of bone in these image sets. This in turn should lead to DRRs more similar to those generated with the real CT.

Looking at the DVHs in Figure 4.4a, it was clear that the largest dosimetric deviation was present in the eyes. It was not possible to visually distinguish the DVHs originating from the model with dUTE+mDixon input and the model with dUTE only input. Looking at the DVH points in Table 4.4, the visually large deviations observed for the eye volume was confirmed by $> 1\%$ deviations in D_{mean} for that volume for both models with and without mDixon input. Inspection of the pCTs of patient h03 revealed that parts of both eyes were given a bone density value, so this may be the reason for the deviations. In general, all deviations were $< 2\%$ and the difference between using mDixon input and dUTE only input was not obvious.

Summary: The reproducibility of results when using a GMR model with dUTE input was demonstrated in this study. From the results obtained, there was an indication that the addition of the mDixon input could improve both the prediction accuracy and the geometric accuracy. The dosimetric accuracy was within 2 % for models both with and without mDixon input.

5.3.2 Random Forest Regression

Looking at the MAPD and MPD curves in Figure 4.3b, it was seen that using the mDixon input improved the prediction accuracy for CT values above 500 HU. In Table 4.2, the MAPDs for both the entire image sets and for voxels inside the air mask were lower when using the mDixon input. The difference was more distinct when looking inside the air mask. As was the case for GMR, the pCTs predicted with RaFR using mDixon input had more noise in the air region which may be what caused this behaviour.

The geometric evaluation showed the same tendency as the prediction deviations. Using the mDixon input resulted in consistently lower bone miss volumes and false bone volumes and a higher Dice score compared to not using it.

As with GMR, the DVHs showed the largest deviations in the eye volume, which was confirmed by the D_{mean} DVH points in Table 4.4. Again, the pCTs revealed that some voxels in the eyes were given a bone density value, which may be the reason for the deviations. It was hard to spot a clear tendency from the DVH points in favour of using mDixon input. The deviations were $< 1.7\%$ for all DVH points.

Summary: The impact of using mDixon image sets as an additional input to a RaFR model was demonstrated. It was found that using mDixon input improved the prediction accuracy and geometric accuracy of the models. It was not possible to conclude whether the dosimetric accuracy was improved when adding mDixon input.

5.3.3 Comparing GMR and RaFR

In Figure 4.5a, the MAPD and MPD curves were shown for GMR and RaFR models with mDixon input. For high density values, the RaFR model had the lowest MAPD. In Figure 4.5b, the same curves were shown for models with dUTE only input with an indication that the GMR model performed better in the high density region.

Looking at the MAPD inside the air mask in Table 4.2, RaFR using mDixon input consistently produced the lowest error. On average, the next best model was GMR with mDixon input followed by RaFR with dUTE only input and lastly GMR with dUTE only input.

From a geometrical viewpoint, RaFR with dUTE+mDixon input produced the best results. The worst results were obtained with GMR using dUTE only input. Which one of GMR with dUTE+mDixon input and RaFR with dUTE only input performed the best was not clear.

Looking at the DVHs and DVH points, it was hard to decide which model to prefer. The largest deviation from the CT based dose was 1.83 % and the difference in the individual deviations was a maximum of 0.38 percentage points. It should be investigated whether these numbers are significant. For the eye volume all pCTs showed the largest dosimetric deviation from the real CT. As already touched upon, prediction errors were present in the eyes causing some voxels to appear as bone. This is probably what caused the dose distribution to appear less homogeneous in this volume.

When looking at prediction deviations in the air mask and the geometrical accuracy, the results were mostly in favour of using RaFR with mDixon input compared to any GMR model and RaFR with dUTE only input. This means that the pCTs produced with this model would probably produce the most satisfactory DRRs. The next best models were GMR with dUTE+mDixon input and RaFR with dUTE only input. Because the dosimetric evaluation was less conclusive, the question arises whether the improvement in geometrical accuracy is worth the extra scan time that the mDixon sequence requires (approximately 2.5 minutes). It could be that the geometrical accuracy of one of the dUTE only models is sufficient for creating usable DRRs. The RaFR model with dUTE only input had an almost similar Dice score to the GMR model with dUTE+mDixon so it may be worth investigating if this model produces sufficient pCTs for RT planning.

Another thing to take into account is the time efficiency of the two regression implementations. On a Linux node with 2×2.66 GHz quad-core processors and 24 GB memory, the current implementations of GMR and RaFR used on average 50 hours to train a model using dUTE+mDixon data from two patients. When using dUTE only data the time was reduced to approximately 35 hours. To predict a new pCT took approximately 20 minutes for GMR and 5 hours for RaFR. These numbers speak in favour of using GMR.

It should also be noted that it was not investigated whether 20 components used in the GMM was the optimal number suited for the task. The 20 components were chosen based on what Johansson *et al.* used in their study. It may be

the case that when adding the mDixon input, more components are needed to model the joint distribution of MRI and CT intensities. A different number of components could potentially improve the prediction performance of GMR.

5.3.4 General Comments on the pCTs

In Figure 4.6, axial slices of all the predicted pCTs and the real CT for patient h03 were shown. As already mentioned, the pCTs predicted with models using mDixon input showed errors in the air region surrounding the head. This is most clearly observed in the inferior axial slices shown in the top row of Figure 4.6. Another interesting thing to note in this row is that the streak artefact present in the real CT is not predicted in the pCTs. This could be seen as an advantage because it would mean that such artefacts would not be present in the DRRs generated from pCTs. It should be noted that streak artefacts were present in all the CTs used in this project. Except for patient h02, these were not removed from the training data. Optimally, the models should be trained on images without such artefacts, because they would then be better representations of the general appearance of a CT image. The pCTs were compared to the real CT when calculating the piecewise deviations. This included the region where streak artefacts were present and this could be the reason that for high density values in the real CT the pCT has a large MAPD and MPD.

From inspection of the pCTs, it was observed that the regions with the largest visual deviations from the real CT were found near the nose, ears and paranasal sinuses. These are all regions with air-tissue interfaces so a reasonable explanation could be that susceptibility artefacts in the MRI image sets are causing this.

Using dUTE+mDixon input produced the lowest overall error inside the air mask for both GMR and RaFR, so it seemed that the mDixon image sets included some information that was not captured by the dUTE image sets. As previously mentioned, the reason for using the mDixon water and fat image sets, was to compensate for water/fat phase cancellation artefacts in the dUTE image sets recorded on a 1 T scanner. These artefacts could cause non-bone voxels to appear as bone. In other words the mDixon input should improve the prediction accuracy of non-bone voxels. The largest difference in the MAPD and MPD curves was observed in the bone region of the real CT which was not as expected. No reasonable explanation to this behaviour was found during this project.

5.4 Evaluation Methods

In this section some of the limitations of the evaluation methods will be outlined. For all methods used in this report, one limitation is that the spatial information is lost. This means that the methods cannot reveal in what spatial region of the image set an error is present. This information would be useful since it may show in what regions the input MRI image sets lack information or are influenced by artefacts. From visual inspection of the pCTs, it seems that errors are made in the paranasal sinuses, nose and ear regions. None of the evaluation methods can back this up. The idea of calculating the MAPD for local regions was carried out to some extend by using the air mask as a region, but it may be useful to look into calculating errors in smaller regions in order to quantify the local errors better. This could also be done for the geometric evaluation. A subtraction of the reference CT and pCT could also serve to show in which spatial regions the errors are most profound.

5.4.1 Prediction Deviation Method

The MAPD and MPD computed in bins of 20 HU provide a tool for quantifying the errors made in specific voxel intensity ranges of the real CT. Taking advantage of the Hounsfield scale, it is thus possible to quantify the errors made in specific tissues. To calculate the error at -500 HU, all voxels in the real CT with an intensity of -510 HU to -490 HU along with the spatially corresponding voxels in the pCT are used. This fact must be kept in mind when inspecting the MAPD and MPD curves, because, depending on the the number of voxels in a bin, the error may be calculated based on 1 voxel or thousands of voxels. If it is calculated based on a proportionally small number of voxels, a seemingly large error in a specific range may not affect the overall image much. This property of the error curves does not mean that they cannot be used to compare and elucidate the models' performance in the different tissue regions, it means that they may not provide a measure of the overall prediction accuracy. To quantify the overall prediction accuracy of the models, the MAPDs based on the entire image set and on the region within the air mask have been calculated.

5.4.2 Geometric Evaluation Method

Even though all images were co-registered prior to training the models, the registration was lost when the pCTs were imported into Varian Eclipse Treatment Planning System for geometric evaluation. For this reason a new registration

had to be made. The geometrical agreement, as measured by the Dice coefficient, may include registration errors since it is a measure of the overlap between regions, which is affected by the registration. However, the same registration was applied to all pCTs belonging to a given patient, so if a registration error was present, the same error would be incorporated into all Dice scores of the same patient. This means that the intra-patient Dice scores should be comparable, but one should keep in mind that when comparing inter-patient scores the difference may be attributed to different quality of the registration.

5.4.3 Dosimetric Evaluation Method

The main limitation when using dose-volume histograms (DVHs) and DVH points for dosimetric evaluation is that even though each DVH is confined to specific regions (PTV or PRVs), the spatial information within these regions is lost. This means that if the simulated dose distributions in each volume are similar using CT and pCT, spatially local deviations in the estimated delivered dose may still exist. Comparing isodose contours in the CT and pCT could provide a spatial evaluation of the dosimetric deviations.

5.5 Future Work

One thing that can affect the quality of the pCT is the amount of data used to train the regression models. The assumption when estimating the model parameters from training data is that the data is representative of all future inputs and outputs. If this is not the case, i.e. the training data represents a rarely observed case of all possible inputs and outputs, the resulting model will not be able to make good predictions on future inputs. Artefacts in the MRI and CT image sets, such as those originating from dental fillings, are examples of patient specific deviations from the general appearance of input and output data. Using data from many patients to train the models increases the probability that the majority of the training data is representative of future inputs and outputs and should in turn lead to better predictions.
In this report, the LBP input study was carried out by using data from one patient and then predicting the pCT of another. The results presented in the mDixon input studies were based on data from three patients using a leave-one-out approach. Adding more patient data to all these studies could improve the overall predictions and may help to determine whether the observed trends are consistent.

When doing regression using GMR, it is possible to get an estimate of the conditional variance of the predictions. This is also possible for RaFR, if using the method of Criminisi *et al.* The variance could be used to show in which regions of the pCT the predictions have a large uncertainty and in which they do not. Along the same line, Johansson *et al.* suggested that the expected absolute deviation could be incorporated into the absorbed dose calculated using the pCT [16]. This would then give an estimate of the uncertainty of the calculated dose, which could be useful as a quality assurance measure. In this project, the conditional variance of the predictions was not investigated, so this is something to look into in the future.

A number of improvements could be made to the LBP implementation used in this project. The LBP approach was inspired by the work by Pauly *et al.* who used a 3-scale 3D LBP-like feature derived from Dixon MRI image sets to predict organ bounding boxes [17]. In their description of the feature extraction, no details as to the radius of the spheres defining the neighbourhood or the size of the 3D cubes was found. These two parameters are of vital importance to the appearance of the resulting feature images, since they define the scale of the neighbourhood that is captured. In this project, it was not investigated whether the three scales used for feature extraction were the optimal for capturing textural information in the MRI image sets. Furthermore, because the LBP contains neighbourhood information, the rotation of the image sets affects the result. The current implementation is not rotation invariant, so this could be included to improve the robustness.

From an time-efficiency point of view, especially the RaFR implementation has room for improvements. The built in Matlab implementation was run in serial, meaning that one tree was trained at a time. The RaF framework is an obvious candidate for parallel computing because each tree is independent from the others. This could improve the speed of both training and predicting. It was not possible to get the built in Matlab implementation to run in parallel, but our own implementation, used to generate the examples in Section 2.5, was successfully run in parallel with marked improvements in speed as a result.

Finally, a clinical validation of the pCTs could be carried out in the future. This would include investigating if the geometrical accuracy is good enough to generate satisfactory DRRs for patient alignment and investigating if the dosimetric deviations are significant compared to a CT based plan.

Chapter 6

Conclusion

In this project, Gaussian mixture regression (GMR) and Random Forest regression (RaFR) were investigated as a tool for obtaining pseudo-CTs (pCTs) from MRI image sets to use in MRI-only radiation therapy planning. dUTE and mDixon MRI image sets were investigated as input to both models and for RaFR a Local Binary Pattern (LBP) feature input was investigated. The results were evaluated in terms of the voxel-wise deviations described by the mean absolute prediction deviation (MAPD) and mean prediction deviation (MPD). A geometrical evaluation of bone in the pCTs was carried out using the Dice coefficient and the missed bone and false bone percentages. Furthermore, a dosimetric evaluation was carried out by comparing dose-volume histograms (DVHs) and DVH points based on pCT and CT dose calculations.

Gaussian Mixture Regression

GMR is a promising approach to generating pCTs. In this project the reproducibility of previously reported results was demonstrated in terms of a similar average MAPD. A new approach of adding mDixon image sets to the input of the model showed to improve the prediction accuracy and geometrical accuracy of the predicted pCTs compared to using only dUTE image input. An average Dice score of 0.76 for dUTE+mDixon models and 0.72 for dUTE only

models was obtained. In terms of the dosimetric accuracy, it was found that the absorbed dose calculations based on GMR pCTs had deviations of $< 2\%$ in clinically relevant DVH points compared to dose calculations based on CT. No clear dosimetric difference was observed between adding the mDixon input and not adding it.

Random Forest Regression

RaFR is another promising method which has not been used to generate pCTs prior to this project. For the implementation of RaFR used in this project, it was found that using a minimum leaf size of 10 as stopping criteria for tree growing resulted in the smallest mean squared prediction deviation (MSPD).
It was investigated whether adding an LBP feature image as input could improve the pCTs. It was found that a dUTE+LBP input improved the prediction accuracy compared to using only dUTE input. The results were based on a relatively small amount of data so further studies should be carried out. The impact of adding mDixon input was also investigated for RaFR. Here it was found that using dUTE+mDixon input improved the prediction accuracy and geometrical accuracy of the pCT compared to using only dUTE input. An average Dice score of 0.80 for dUTE+mDixon models and 0.75 for dUTE only models was obtained. The dosimetric deviations between absorbed dose calculations based on RaFR pCTs and real CT were found to be $< 1.7\%$ in clinically relevant DVH points.

Comparison of Gaussian Mixture Regression and Random Forest Regression

The RaFR model with dUTE+mDixon input produced the best pCTs in terms of prediction accuracy and geometrical accuracy. The next best models were GMR with dUTE+mDixon input and RaFR with dUTE only input. The extra scan time required with the mDixon input and the time efficiency of the GMR and RaFR implementations should also be considered, before drawing conclusions as to which model is preferred.

Future Work

The results presented in this report showed the potential of using GMR and RaFR for predicting pCTs. A number of things could be investigated in the

future. First of all, studies using more data could be done in order to verify the trends observed in this report. An investigation of the significance of the geometric and dosimetric deviations could be carried out. This would also help in determining if the improvement in geometrical accuracy observed with the mDixon input is worth the extra scan time.

The LBP input study showed promising results but further investigation is needed. Using feature images with neighbourhood information could prove to make additional non-conventional MRI sequences like the mDixon redundant.

Appendix A

Abstract Accepted for ESTRO Forum 2013

The impact of a Dixon sequence in creating a pseudo CT scan from MR images using a Gaussian mixture regression model

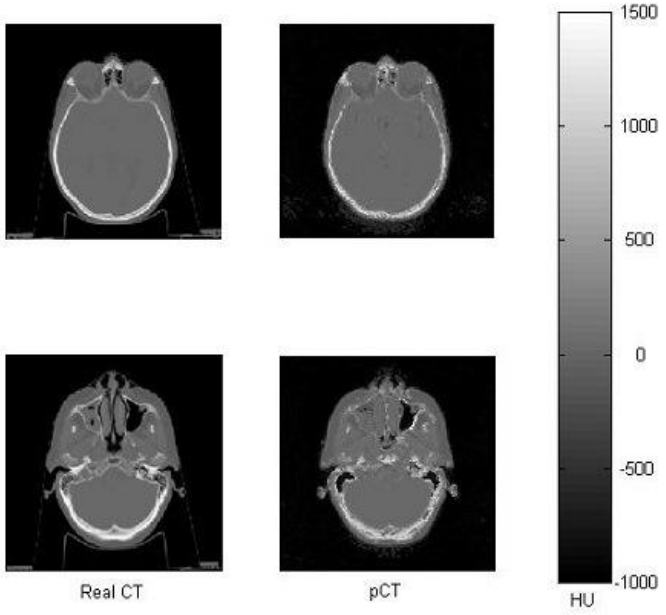
D. Andreasen[1], J.L. Andersen[2], R.H. Hansen[3], K. Van Leemput[1], J.M. Edmund[2]

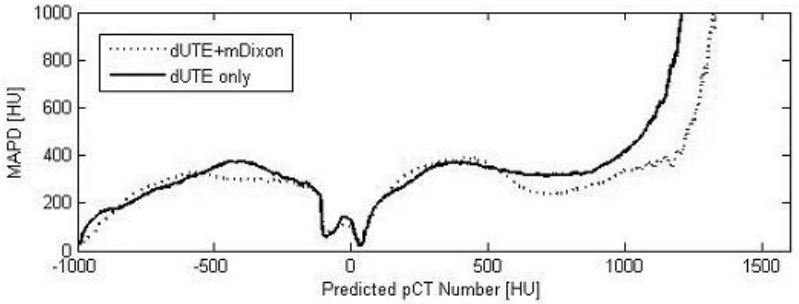
[1]Technical University of Denmark, Department of Informatics and Mathematical Modelling, Lyngby, Denmark.
[2] Copenhagen University Hospital, Herlev, Department of Oncology, Herlev, Denmark.
[3]Copenhagen University Hospital, Herlev, Department of Radiology, Herlev, Denmark

Purpose: For RT based on MRI only, a promising approach is to obtain a substitute CT scan from the MR images (a so-called pseudo CT, pCT) using a Gaussian mixture regression (GMR) model. The GMR model has previously been investigated on 3T MR images using a dual ultra-short echo time (dUTE) sequence and was shown to give sufficient information for training the GMR model. The dUTE sequence provides contrast between bone and tissue using dual echo times but at 1 T, chemical shift artifacts at the second echo time may cause voxels containing water and fat to behave like bone. The multiecho-Dixon (mDixon) MR sequence provides contrast between water and fat which could potentially remove this problem and provide valuable information for the GMR model. In this study, we investigate the robustness of the GMR model on predicting pCT scans from dUTE MR images of a 1 T scanner and how adding an mDixon sequence affects the generated pCT.

Methods: Head scans of 2 patients fixated for whole brain RT were acquired on a 1 T open MR scanner with flex coils. dUTE sequences were obtained at flip angles 10 and 25 degrees, respectively. Echo- and repetition times TE1/TE2/TR were 0.09/3.5 /7.1 ms with a voxel resolution of 1x1x1 mm and a 256 mm FOV. The mDixon was acquired with TE1/dTE/TR equal to 6.9/3.5/16 ms, a voxel resolution of 1x1x1.5 mm and a 250.5 mm FOV. CT head scans were acquired with a voxel resolution of 0.6x0.6x2 mm and a 220 mm FOV. The CT was registered to the high angle TE1 UTE using a mutual information algorithm and all MR scans were internally registered. All scan were resliced to the dUTE resolution and cropped to the smallest FOV. The MR images were low- and high-pass filtered creating two new images per filtered image. The MR images, their filtered counterparts and the CT image were considered as random variables and the voxel intensities a sample from their underlying distribution. A GMR model was initialized with 20 centers using k-means clustering and an EM algorithm was used to train the model on the data from one of the patients. The model was then applied on the other patient to generate the pCT. A model using only the dUTE images and one adding the mDixon images were trained. A comparison using the real CT to calculate the mean absolute prediction error (MAPD) of the pCT in bins of 20 HU was carried out.

Results: The pCTs of one patient using the extended model is shown in the figure. Qualitatively (upper images) and quantitatively (lower graph), the results are similar to those previously reported for 3T using dUTE only. A reduction in MAPD can be observed in the bone region (>500 HU) by adding mDixon to the model.





Top: Real CT (left) and corresponding pseudo CT (right) for a superior (upper) and inferior (lower) axial slice of the head. A HU grayscale is indicated on the far right. **Bottom:** The MAPD vs. pseudo CT in bins of 20 HU for the dUTE (solid) and the dUTE+mDixon (dotted) model.

Conclusion: The robustness of a GMR model on 1T MR images was demonstrated. The model was further expanded with an mDixon sequence which reduced the prediction error of predicted CT values >500 HU. Although a study based on larger amounts of data should be carried out, there is an indication that the mDixon sequence improves CT prediction from dUTE MR images.

List of Figures

List of Tables

Bibliography

[1] J. Purdy, J. Michalski, J. Bradley, S. Vijayakumar, C. Perez, and S. Levitt, "Three-dimensional treatment planning and conformal therapy," in *Technical Basis of Radiation Therapy* (S. Levitt and J. Purdy, eds.), Medical Radiology, pp. 179–202, Springer Berlin Heidelberg, 2006.

[2] V. S. Khoo and D. L. Joon, "New Developments in MRI for Target Volume Delineation in Radiotherapy," *The British Journal of Radiology*, vol. 79, pp. S2–S15, 2006.

[3] B. H. Kristensen, F. J. Laursen, V. Logager, P. F. Geertsen, and A. Krarup-Hansen, "Dosimetric and geometric evaluation of an open low-field magnetic resonance simulator for radiotherapy treatment planning of brain tumours," *RADIOTHERAPY AND ONCOLOGY*, vol. 87, pp. 100–109, APR 2008.

[4] T. Nyholm, M. Nyberg, M. Karlsson, and M. Karlsson, "Systematisation of spatial uncertainties for comparison between a mr and a ct-based radiotherapy workflow for prostate treatments," *Radiation Oncology*, vol. 4, no. 1, p. 54, 2009.

[5] A. Johansson, M. Karlsson, and T. Nyholm, "CT Substitute Derived from MR Sequences with Ultrashort Echo Time," *Medical Physics*, vol. 38, no. 5, pp. 2708–2714, 2011.

[6] T. Boettger, T. Nyholm, M. Karlsson, C. Nunna, and J. C. Celia, "Radiation Therapy Planning and Simulation with Magnetic Resonance Images," *Proc. of SPIE*, 2008.

[7] V. Keereman, P. Mollet, Y. Berker, V. Schulz, and S. Vandenberghe, "Challenges and Current Methods for Attenuation Correction in PET/MR," *Magn Reson Mater Phy*, 2012.

[8] M. D. Robson, P. D. Gatehouse, M. Bydder, and G. M. Bydder, "Magnetic Resonance: An Introduction to Ultrashort TE (UTE) Imaging," *Journal of Computer Assisted Tomography*, vol. 27, no. 6, pp. 825–846, 2003.

[9] Y. Berker, J. Franke, A. Salomon, M. Palmowski, H. C. Donker, Y. Temur, F. M. Mottaghy, C. Kuhl, D. Izquierdo-Garcia, Z. A. Fayad, F. Kiessling, and V. Schulz, "MRI-Based Attenuation Correction for Hybrid PET/MRI Systems: A 4-Class Tissue Segmentation Technique Using a Combined Ultrashort-Echo-Time/Dixon MRI Sequence," *Journal of Nuclear Medicine*, vol. 53, no. 5, pp. 796–804, 2012.

[10] V. Keereman, Y. Fierens, T. Broux, Y. D. Deene, M. Lonneux, and S. Vandenberghe, "MRI-Based Attenuation Correction for PET/MRI Using Ultrashort Echo Time Sequences," *J Nucl Med*, vol. 51, pp. 812–818, 2010.

[11] C. Catana, A. van der Kouwe, T. Benner, C. J. Michel, M. Hamm, M. Fenche, B. Fischl, B. Rosen, M. Schmand, and A. G. Sorensen, "Toward Implementing an MRI-Based PET Attenuation-Correction Method for Neurologic Studies on the MR-PET Brain Prototype," *J Nucl Med*, vol. 51, 2010.

[12] H. M. Kjer, "MRI in Radiation Therapy Planning: The Use of Ultrashort Echo Time Imaging," Master's thesis, DTU, 2012.

[13] A. Martinez-Möller, M. Souvatzoglou, G. Delso, R. A. Bundschuh, C. Chefd'hotel, S. I. Ziegler, N. Navab, M. Schwaiger, and S. G. Nekolla, "Tissue Classification as a Potential Approach for Attenuation Correction in Whole-Body PET/MRI: Evaluation with PET/CT Data," *J Nucl Med*, vol. 50, pp. 520–526, 2009.

[14] M. Eiber, A. Martinez-Möller, M. Souvatzoglou, K. Holzapfel, A. Pickhard, D. Löffelbein, I. Santi, E. J. Rummeny, S. Ziegler, M. Schwaiger, S. G. Nekolla, and A. J. Beer, "Value of a Dixon-based MR/PET attenuation correction sequence for the localization and evaluation of PET-positive lesions," *Eur J Nucl Med Mol Imaging*, vol. 38, pp. 1691—-1701, 2011.

[15] M. Hofmann, F. Steinke, V. Scheel, G. Charpiat, J. Farquhar, P. Aschoff, M. Brady, B. Schölkopf, and B. J. Pichler, "MRI-Based Attenuation Correction for PET/MRI: A Novel Approach Combining Pattern Recognition and Atlas Registration," *Journal of Nuclear Medicine*, vol. 49, pp. 1875–1883, 2008.

[16] A. Johansson, M. Karlsson, J. Yu, T. Asklund, and T. Nyholm, "Voxel-wise uncertainty in CT substitute derived from MRI," *Med. Phys.*, vol. 39, no. 6, pp. 3283–3290, 2012.

[17] O. Pauly, B. Glocker, A. Criminisi, D. Mateus, A. Möller, S. Nekolla, and N. Navab, "Fast multiple organ detection and localization in whole-body mr dixon sequences," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011* (G. Fichtinger, A. Martel, and T. Peters, eds.), vol. 6893 of *Lecture Notes in Computer Science*, pp. 239–247, Springer Berlin Heidelberg, 2011.

[18] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu, "Regression forests for efficient anatomy detection and localization in ct studies," in *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging* (B. Menze, G. Langs, Z. Tu, and A. Criminisi, eds.), vol. 6533 of *Lecture Notes in Computer Science*, pp. 106–117, Springer Berlin Heidelberg, 2011.

[19] D. J. Tyler, M. D. Robson, R. M. Henkelman, I. R. Young, and G. M. Bydder, "Magnetic Resonance Imaging With Ultrashort TE (UTE) PULSE Sequences: Technical Considerations," *Journal of Magnetic Resonance Imaging*, vol. 25, pp. 279–289, 2007.

[20] J. Rahmer, P. Börnert, J. Groen, and C. Bos, "Three-Dimensional Radial Ultrashort Echo-Time Imaging with T2 Adapted Sampling," *Magnetic Resonance in Medicine*, vol. 55, pp. 1075–1082, 2006.

[21] W. T. Dixon, "Simple proton spectroscopic imaging.," *Radiology*, vol. 153, no. 1, pp. 189–194, 1984.

[22] E. Delfault, J. Beltran, G. Johnson, J. Rousseau, X. Marchandise, and A. Cotten, "Fat Suppresion in MR Imaging: Techniques and Pitfalls," *RadioGraphics*, 1999.

[23] J. Ma, "Dixon Techniques for Water and Fat Imaging," *Journal of Magnetic Resonance Imaging*, vol. 28, pp. 543–558, 2008.

[24] T. Perkins, J. Van Tilburg, G. Herigault, H. Eggers, A. Duijndam, G. Beck, and S. Hussain, "Preliminary Clinical Experience with a Multiecho 2-Point DIXON (mDIXON) Sequence at 3T as an Efficient Alternative for Both the SAR-intensive Acquired In- and Out-of-Phase Chemical Shift Imaging as well as for 3D Fat-suppressed T1-weighted Sequences used for Dynamic Gadolinium-enhanced Imaging," *Proc. Intl. Soc. Mag. Reson. Med.*, vol. 18, p. 556, 2010.

[25] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Merchal, "Automated multi-modality image registration based on information theory," *Information Processing in Medical Imaging*, vol. 3, pp. 263–274, 1995.

[26] J. Pluim, J. Maintz, and M. Viergever, "Mutual-information-based registration of medical images: a survey," *Medical Imaging, IEEE Transactions on*, vol. 22, pp. 986 –1004, aug. 2003.

[27] D. Cohn, Z. Ghahramani, and M. Jordan, "Active Learning with Statistical Models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[28] K. Van Leemput and R. Larsen, *02505 Course Note Medical Image Analysis*. DTU Informatics, 2011.

[29] H. Sung, *Gaussian Mixture Regression and Classification*. PhD thesis, Rice University, 2004.

[30] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classificcation and Regression Trees*. Wadsworth, 1984.

[31] L. Breiman, "Random Forests," *Machine Learning*, 2001.

[32] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2–3, pp. 81–227, 2011.

[33] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *PATTERN RECOGNITION*, vol. 29, pp. 51–59, JAN 1996.

[34] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 971 –987, jul 2002.

[35] J. A. Dowling, J. Lambert, J. Parker, O. Salvado, J. Fripp, A. Capp, C. Wratten, J. W. Denham, and P. B. Greer, "An Atlas-Based Electron Density Mapping Method for Magnetic Resonance Imaging (MRI)-Alone Treatment Planning and Adaptive MRI-Based Prostate Radiation Therapy," *Int J Radiation Oncol Biol Phys*, vol. 83, no. 1, pp. e5–e11, 2012.

[36] "3. special considerations regarding absorbed-dose and dose–volume prescribing and reporting in imrt," *Journal of the ICRU*, vol. 10, no. 1, pp. 27–40, 2010.

[37] A. S. Glassner, ed., *Graphics Gems*. Academic Press, 1990.

[38] S. Calinon, F. Guenter, and A. Billard, "On Learning, Representing and Generalizing a Task in a Humanoid Robot," *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 37, no. 2, pp. 286–298, 2007.

[39] M. Hofmann, B. Pichler, B. Schölkopf, and T. Beyer, "Towards Quantitative PET/MRI: a Review of MR-based Attenuation Correction Techniques," *Eur J Nucl Med Mol Imaging*, vol. 36, pp. 93–104, 2009.