Combining nonlinear least squares and random forest regression to increase the accuracy and precision of DCE-MRI tracer kinetic model parameter estimates

J. Palowski[1,2], and C. J. Rose[1,2]

[1]The University of Manchester Biomedical Imaging Institute, The University of Manchester, Manchester, Greater Manchester, United Kingdom, [2]Manchester Academic Health Science Centre, The University of Manchester, Manchester, Greater Manchester, United Kingdom

INTRODUCTION Nonlinear regression is commonly used in MRI to obtain point estimates of physiological quantities that cannot be measured directly, but can be modelled as a function of measureable phenomena: examples include the spin-lattice and spin-spin relaxation times T_1 and T_2, and microvascular characteristics arising from tracer kinetic modelling[1]. Nonlinear regression is often posed as an optimisation problem in which the model parameters are manipulated to minimise some measure of dissimilarity—typically the sum of squared differences—between observed data and the model's prediction of the data. Conceptually, regression is just a mapping, R, between the space in which measurements are made, X, and the space of the quantity of ultimate interest, Y—i.e., $R: X \rightarrow Y$, $y = R(\boldsymbol{x})$. Aside from least squares, other regression methods exist and have been used to estimate physiological quantities in MRI, such as Bayesian maximum *a posteriori* estimation[2]. We have developed and evaluated a method to estimate the parameters of the extended Tofts version of the Kety model[1]—as applied in dynamic contrast-enhanced MRI—based on a machine learning method called random forests[3]. Using simulated gadopentate dimeglumine (Gd-DTPA) concentration time series we show that, compared to conventional least squares, the proposed method can estimate all three parameters of the model with greater accuracy and comparable precision, and in less time.

RANDOM FORESTS Random forests comprise many regression trees. A regression tree is a piecewise constant approximation of a function of interest: the approximation is learned from training data in the form of $(\boldsymbol{x}_i, y_i)$ pairs, for i=1…N. Trees are so-called because they can be represented graphically as a tree structure: to estimate the y value associated with an observation $\boldsymbol{x}$, a decision is made at each node of the tree—starting with the top-most—on the basis of the predictor variables (i.e., elements of the observed vector $\boldsymbol{x}$), until a terminal node is reached; the value associated with that node is the tree's estimate of the y value. Fig. 1 shows an example tree in which a function of two variables is approximated. Methods exist to learn regression trees (i.e., their structure, node variables and thresholds). Unfortunately, the piecewise constant approximations limit trees' predictive efficacy; this can be improved by increasing the number of terminal nodes, but this often leads to over-fitting on the training data (poor generalisation to unseen data). Random forests overcome this problem by using a large number of regression trees, each trained on a bootstrapped sample of the training data (see Ref. 3 for details). Given an observation $\boldsymbol{x}$, a forest's estimate of y is the average of each tree's estimate. Random forests can approximate highly non-linear relationships with high accuracy and precision. Because each tree's estimate can be computed quickly—by comparing elements of an observation $\boldsymbol{x}$ to a set of thresholds—a random forest estimate can also be computed quickly. Random forests also generate unbiased internal estimates both of generalisation error and how important each variable (element of $\boldsymbol{x}$) is to the estimation task. Random forests may be extended to give distributions or confidence intervals on parameter estimates.

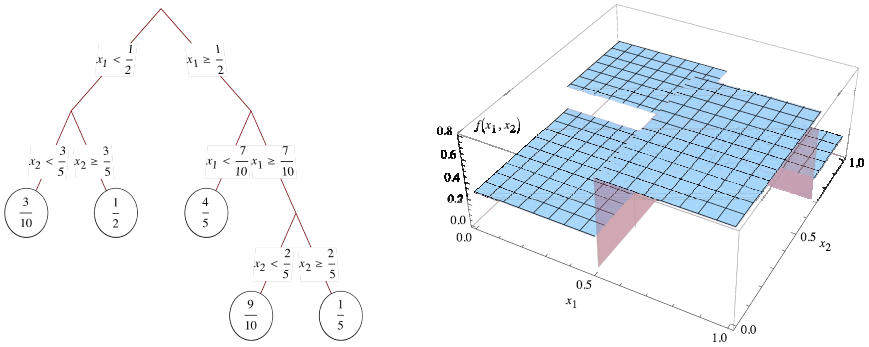

Fig. 1 *A single example regression tree, approximating a function $f(x_1, x_2)$ as a piecewise constant function.* Left: *At each node, the tree splits in two on the basis of a predictor variable; terminal nodes show the approximated value of the function f.* Right: *A plot of the regression tree's approximation of the function f, illustrating how trees subdivide the space of predictor variables into rectangles (or orthotopes in higher dimensions) and map these to constant values.*

METHOD The extended Tofts version of the Kety model explains observed contrast agent concentration in terms of the transfer coefficient K^{trans} (units: min^{-1}) and the relative volumes of the blood plasma and extravascular extracellular spaces, v_p and v_e respectively, as:

$$C_t(t) = v_p\, C_p(t) + K^{\text{trans}} \int_o^t C_p(t')\, e^{-\frac{K^{\text{trans}}}{v_e}}\, dt' \qquad (1)$$

where $C_t(t)$ is the contrast agent concentration (in mmol) at time t (in min) and $C_p(t)$ is the contrast agent concentration in the arterial blood plasma at time t. Conventionally, the parameters K^{trans}, v_p and v_e are estimated by manipulating them such that the sum of squared differences between the predicted and observed $C_t(t)$ values is minimised. We wish to estimate them using random forest regression. We simulated two sets of 1000 $C_t(t)$ time series (using a population averaged[4] $C_p(t)$) — using acquisition timings from a typical DCE-MRI protocol—by randomly sampling "known" parameters (K^{trans}, v_p, v_e) uniformly such that $0 \leq v_p+v_e \leq 1$ and $0 \leq K^{trans} \leq 1.4$; realistic levels of noise were then added. Three random forests were trained on the first set of $C_t(t)$ time series to estimate K^{trans}, v_p and v_e (i.e., each forest estimates one of the model parameters). The observed vectors $\boldsymbol{x}_i$ each comprised the contrast agent concentration values ordered by acquisition time: x_i=$C_t(t_i)$, i=1…t_{end}. (The index into $\boldsymbol{x}$ implicitly encodes acquisition time, t, so these times were not included as variables in $\boldsymbol{x}$.) We found that while random forests could estimate v_e and v_p, this was not so for K^{trans}. Our random forest-based algorithm estimates the three parameters as follows: random forests are used to estimate v_e and v_p and these estimates are then used as constraints within a conventional least-squares estimation of K^{trans}, creating a hybrid approach to estimation. Using the second (independent) set of 1000 time series, we compared the hybrid random forest-based method to the conventional least-squares method for estimating all three parameters. We assessed estimation accuracy and precision respectively using the mean and variance of the absolute differences between the known and estimated parameter values. The null hypotheses of no difference in accuracy and precision were tested using t- and F-tests, respectively. Finally, we calculated the average amount of time required to estimate the three parameters for one time series. We performed least-squares regression using Matlab's `lsqcurvefit` (The Mathworks, Natick, MA) and used the freely-available Matlab port of the random forest algorithm[5].

RESULTS

		Accuracy Mean(Absolute Differences)	*P*	Precision Var(Absolute Differences)	*P*	Time (s)
Hybrid Random Forest	K^{trans}	0.1837	<0.0005	0.0748	<0.0005	2.8
	v_e	0.0264	0.0087	0.0017	<0.0005	
	v_p	0.0017	<0.0005	<0.0001	<0.0005	
Least Squares	K^{trans}	0.2302		<0.0001		3.8
	v_e	0.0828		0.4650		
	v_p	0.0087		<0.0001		

CONCLUSIONS In simulated data, the hybrid random forest method can estimate all three parameters with greater accuracy and in less time compared to conventional least squares; the precision of the random forest estimates is poorer for K^{trans}, better for v_e, and comparable for v_p. Future work will evaluate the method in clinical data.

REFERENCES 1 Tofts P, J Mag Reson Imag 1997, 7:91–101. **2** Kelm BM et al., IEEE Trans Med Imag 2009, 28:1534–47. **3** Breiman L, Mach Learn 2001, 45:5–32. **4** Parker G et al., Mag Reson Med 2006, 56:993–1000. **5** Available online at http://code.google.com/p/randomforest-matlab/ (accessed October 2010).