# Parameter Estimation of Compartment Models in `SoilR` Using Classical and Bayesian Optimization

Markus Müller* and Carlos A. Sierra†
Max Planck Institute for
Biogeochemistry

October 9, 2013

## Introduction

The objective of this document is to provide examples on how to use `SoilR` in combination with package `FME` to estimate parameter values of soil organic matter decomposition models using observed data. We will not explain `FME` functionality here, but strongly recommend to read the vignette for package `FME` (Soetaert & Petzoldt, 2010). Instead, we focus here on the application to the type of models implemented in `SoilR`.

We present here two examples, one is the parameterization of a two-pool model with a feedback connection scheme applied to a soil incubation experiment. The other example uses observed radiocarbon data from $CO_2$ measurements conducted at Harvard Forest, USA.

## Example 1: A soil incubation experiment

Measurements of evolved $CO_2$ from incubation experiments provide useful data for parameterizing soil organic matter decomopsition models and identify functionally distinct pools (Schadel et al., 2013). We present here data from an incubation experiment in which we measured the evoloved $CO_2$ from a forest soil. The dataset, `eCO2`, is already included in `SoilR` and includes data from two incubation experiments, one with a temperate forest soil and another with a boreal forest soil. First, we load `SoilR` into our `R` session and extract the data from the boreal site into a separate object excluding the column that identifies the sampling site; column names need to be renamed for consistency with `FME`.

```
> library(SoilR)
> library(FME)
> library(MASS)
> library(lattice)
```

*mamueller@bgc-jena.mpg.de
†csierra@bgc-jena.mpg.de

```
> plot(BorealCO2[,1:2], xlab="Days", ylab="Evolved CO2 (mgC g-1 soil)")
> arrows(BorealCO2[,1],BorealCO2[,2]-BorealCO2[,3],BorealCO2[,1],
+        BorealCO2[,2]+BorealCO2[,3],code=3,angle=90,length=0.1)
```

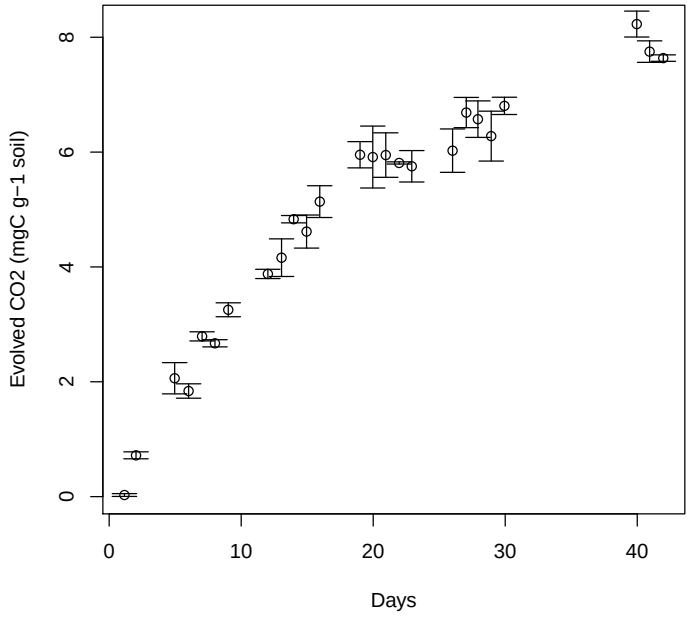


Figure 1: Cummulative evolved $CO_2$ from an incubation experiment with a boreal forest soil.

```
> BorealCO2=subset(eCO2, subset=Sample=="AK_T25", select=-Sample)
> names(BorealCO2)<-c("time","eCO2","eCO2sd")
```

We can plot the data with the command

To this dataset, we are interested in finding parameters for a two-pool model with connection in feedback of the form (Sierra et al., 2012)

$$\begin{aligned}\frac{dC_1}{dt} &= I - k_1 C_1 + \alpha_{1,2} k_2 C_2 \\ \frac{dC_2}{dt} &= \alpha_{2,1} k_1 C_1 - k_2 C_2 \end{aligned} \tag{1}$$

so we are interested in finding the values of the decomposition rates $k_1$ and $k_2$ as well as the transfer coefficient to pool 2 from pool 1 ($\alpha_{2,1}$) and to pool 1 from pool 2 ($\alpha_{1,2}$). Given that the data comes from an incubation experiment, we assume that there are no inputs of carbon, so $I = 0$. In addition, we are also interested in obtaining a value for the partitioning coefficient for the two fractions $\gamma$, so $C_1 = C_{total}\gamma$, and $C_2 = C_{total}(1-\gamma)$.

This model (equation 1) is implemented in `SoilR` with the function `TwopFeedbackModel`. We will find first the best set of parameters that fit the data using

classical optimization using the FME package (Soetaert & Petzoldt, 2010). For this, we need to create a function that takes arbitrary values of the parameters of the model, creates a model in SoilR, calculates the cummulative respiration flux, and returns the output consistent with FME requirements. We also need to create a vector of time steps in days and give the total amount of carbon in the soil at the begining of the experiment (7.7 mg C $g^{-1}$ C).

```
> days=seq(0,42)
> Ctotal=7.7
> eCO2func=function(pars){
+   mod=TwopFeedbackModel(
+   t=days,
+   ks=pars[1:2],
+   a21=pars[3]*pars[1],
+   a12=pars[4]*pars[2],
+   C0=Ctotal*c(pars[5],1-pars[5]),
+   In=0,
+   pass=TRUE
+   )
+   AccR=getAccumulatedRelease(mod)
+   return(data.frame(time=days,eCO2=rowSums(AccR)))
+ }
```

Notice that this function, eCO2func, requires a vector of parameters pars with the values of the two decomposition rates in positions 1 and 2, the values of the two transfer rates in position 3 and 4, and the partitioning coefficient $\gamma$ in position 5. This function returns a data.frame with two columns, time in days and the sum of the cummulative release for the two pools.

The next step is to create a cost function according to FME requirements. This cost function takes as arguments a function with the model, the set of observations, and a measure of the error in the observations. The function calculates sums of squared residuals from the model output and the observed data, which can be further minimized for optimization.

```
> eCO2cost=function(pars){
+   modelOutput=eCO2func(pars)
+   return(modCost(model=modelOutput, obs=BorealCO2, err="eCO2sd"))
+ }
```

This function returns an object of class modCost, which can be further used by FME for local sensitivity analysis, multivariate parameter identifiability, and parameter optimization. We strongly recommend users to to read FME documentation for sensitivity and identifiability analyses. The procedure for optimization consist first on given a set of initial parameter values and then run function modFit for minimizing the cost function. modFit can take as argument upper and lower limits for the parameter values and the optimization method, which for this example we use the Levenberg-Marquardt algorithm.

```
> inipars=c(k1=0.5,k2=0.05,alpha21=0.5,alpha12=0.1,gamma=0.5)
> eCO2fit=modFit(f=eCO2cost,p=inipars,method="Marq",
+                upper=c(Inf,Inf,1,1,1),lower=c(0,0,0,0,0))
```

```
> plot(BorealCO2[,1:2], xlab="Days", ylab="Evolved CO2 (mgC g-1 soil)")
> arrows(BorealCO2[,1],BorealCO2[,2]-BorealCO2[,3],BorealCO2[,1],
+        BorealCO2[,2]+BorealCO2[,3],code=3,angle=90,length=0.1)
> lines(fitmod)
```

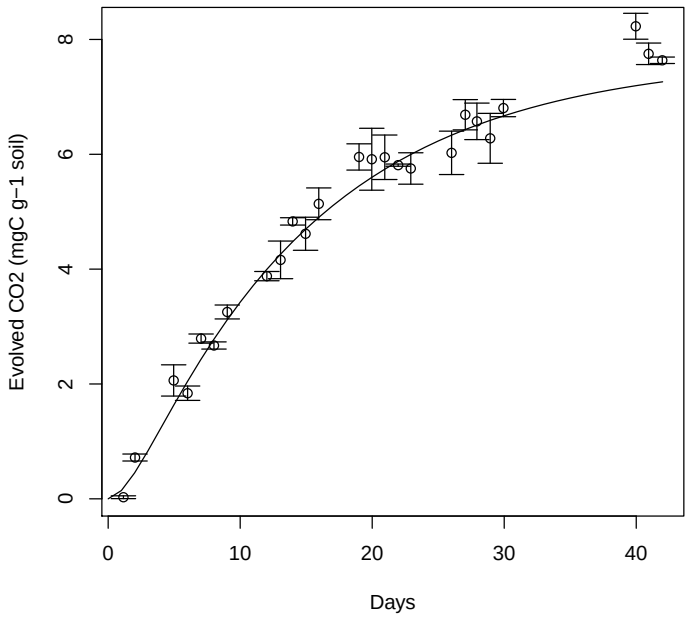


Figure 2: Best fit curve and observed data of CO2 evolved from an incubation experiment.

The best set of parameter values found by the function can be obtained by typing

```
> eCO2fit$par
```

```
      k1        k2   alpha21   alpha12     gamma
0.1826949 0.4754156 0.9930036 0.5219279 0.9944090
```

These set of parameters can be used now to run the model again and plot the obtained model against the observations

```
> fitmod=eCO2func(eCO2fit$par)
```

The results from this optimization can be used for Bayesian parameter estimation with `FME`. For details about the procedure please see Soetaert & Petzoldt (2010). In our example, we need first to extract the variance from the prior optimization and used as the initial variance in the Bayesian procedure and to determine the *jump*, a value that determines how much a new parameter set is deviated from the old one. To avoid long compiling times in `SoilR` we only

```
> pairs(eCO2mcmc)
```

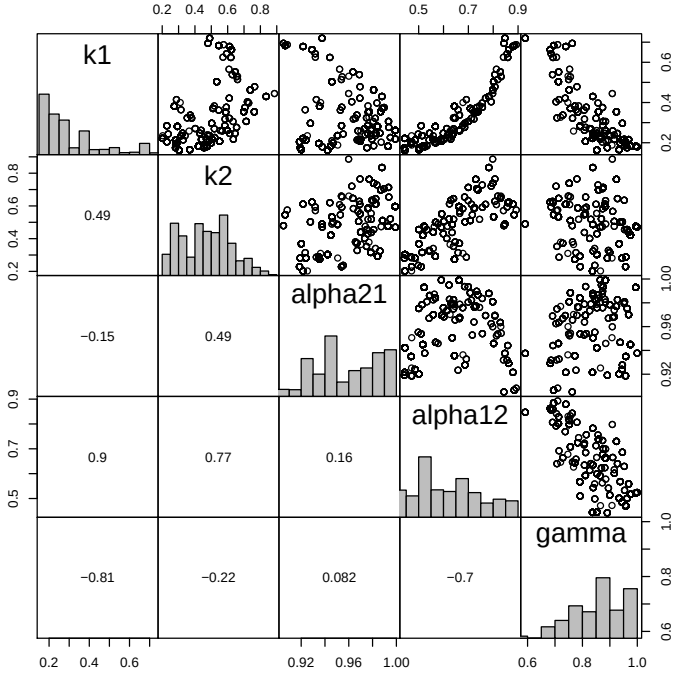


Figure 3: Histogram and scatter plots of the values obtained from the Markov chain Monte Carlo procedure.

use 1000 iterations in this example, but this number can be much larger to guarantee convergence of the chains.

The results of the MCMC procedure can be obtained with the function `summary()`. The output gives the mean, standard deviation, min and max, and 25% quantiles for all parameter values. It also produces summary statistics for the variance of the observed variable.

```
> summary(eCO2mcmc)
```

```
            k1        k2    alpha21   alpha12      gamma   var_eCO2
mean 0.2971989 0.4718510 0.95945984 0.6187460 0.85911836 0.13043344
sd   0.1434748 0.1527655 0.02693437 0.1286495 0.09225655 0.04354060
min  0.1630699 0.2020023 0.90509049 0.4422806 0.59145843 0.06764484
max  0.7197345 0.8861021 0.99916452 0.8927744 0.99946894 0.33667040
q025 0.1832749 0.3157790 0.93770680 0.5167854 0.79056731 0.09631634
q050 0.2475042 0.4754156 0.95444302 0.6015789 0.86474265 0.12631797
q075 0.3540382 0.5962818 0.98415291 0.7124438 0.94415722 0.14190060
```

A plot with the posterior distribution of the obtained parameter values can be obtained with function `pairs` (Figure 3)

For model prediction, it is also possible to use `FME` and function `sensRange` to obtain a graph of the model prediction with uncertainty ranges (Figure 4).

```
> predRange=sensRange(func=eCO2func, parInput=eCO2mcmc$par)
> plot(summary(predRange),ylim=c(0,9),xlab="Days",
+      ylab="Evolved CO2 (mg C g-1 C)",main="")
> points(BorealCO2)
> arrows(BorealCO2[,1],BorealCO2[,2]-BorealCO2[,3],BorealCO2[,1],
+        BorealCO2[,2]+BorealCO2[,3],code=3,angle=90,length=0.1)
```

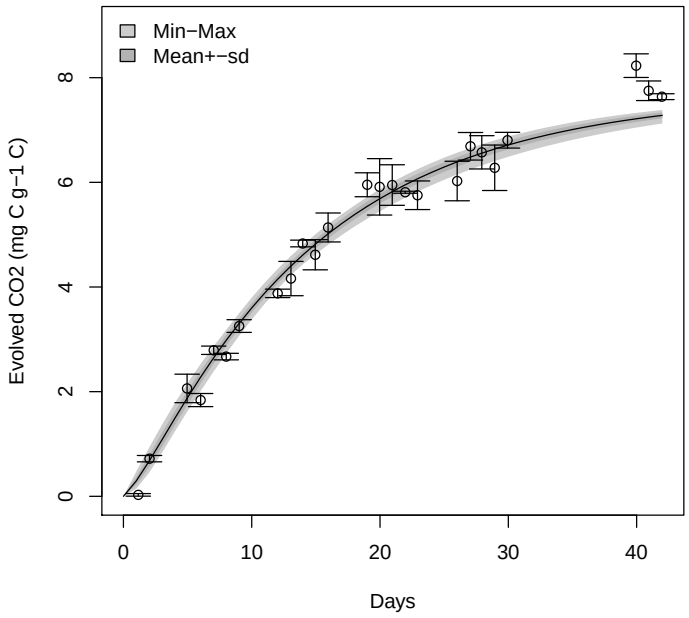


Figure 4: Model predictions using the set of parameters obtained from the MCMC procedure.

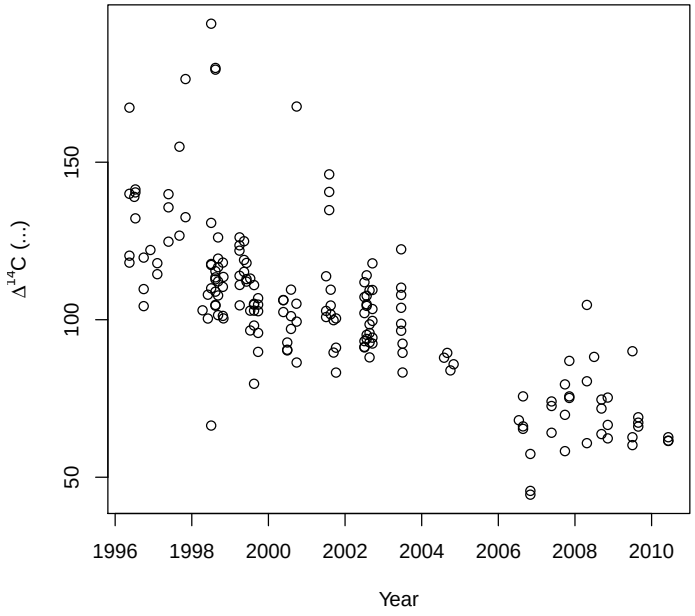


Figure 5: $\Delta^{14}$C value of the respired $CO_2$ in a temperate broadleave forest at Harvard Forest, USA.

It is now obvious from this example that the workhorse of the parameter estimation procedure is the package `FME` of Soetaert & Petzoldt (2010). The main important task to learn about `SoilR` is how to set up the function that runs the model and obtains the variable of interest.

# Example 2: Radiocarbon in respired $CO_2$

`SoilR` can also calculate the amount of radiocarbon in soils or in respired $CO_2$. Here, we take as an example a series of observations of radiocarbon in respired $CO_2$ conducted at Harvard Forest, USA. The dataset is also included in `SoilR`, and can be visialized in Figure 5.

We are interested in fitting the following three-pool model to the data

$$\frac{d\mathbf{C}(t)}{dt} = I \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ 0 \end{pmatrix} + \begin{pmatrix} -k_1 & 0 & 0 \\ a_{21} & -k_2 & 0 \\ a_{31} & 0 & -k_3 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}. \tag{2}$$

where $\gamma_1$ and $\gamma_2$ are known.

For this task, we simply need to prepare a model object in `SoilR` that can be further used by `FME` for parameter estimation. The radiocarbon content of $CO_2$ in the atmosphere is necessary for running the model, because it inform us about the concentration and rate of radiocarbon input to the soil. For this

example we will use the dataset C14Atm_ NH provided with SoilR, but other provided datasets such as Hua2013 can also be used.

First, we define the points in time to run the model from the atmospheric radicarbon dataset

```
> time=C14Atm_NH$YEAR
> t_start=min(time)
> t_end=max(time)
```

To create the vector of input fluxes we need to create a new object of class InputFlux. For our particular model, input fluxes to the $C_1$ and $C_2$ pools are created by this command

```
> inputFluxes=new("InputFlux",
+                 t_start,
+                 t_end,
+                 function(t0){matrix(nrow=3,ncol=1,c(270,150,0))}
+ )
```

assuming that pool 1 receives 270 gC $m^2$ $yr^{-1}$ and pool 2 150 gC $m^2$ $yr^{-1}$.

The initial amount of carbon is created by aggregating the organic and mineral pools for this site reported in Sierra et al. (2012)

```
> C0=c(390,220+390+1376,90+1800+560)
```

We now write a function that creates a Model object in SoilR that takes as arguments a set of parameters and returns the $\Delta^{14}$C value of the respired carbon

The observed data needs to be orginazed in a dataframe of the form

```
> DataR14t=cbind(time=HarvardForest14CO2[,1],
+                R14t=HarvardForest14CO2[,2],
+                sd=sd(HarvardForest14CO2[,2]))
```

With all these elements ready, we can now use FME for the parameter optimization procedure. We will avoid a detailed explanation and present in the following the creation of the cost function, the initial optimization, and the final Bayesian parameter estimation.

The obtained posterior distributions of the parameters can now be assessed graphically (Figure 6). The final model with its uncertainty and how it compares to the data can now be obtained (Figure 7).

```
> pairs(MCMC,nsample=500)
```

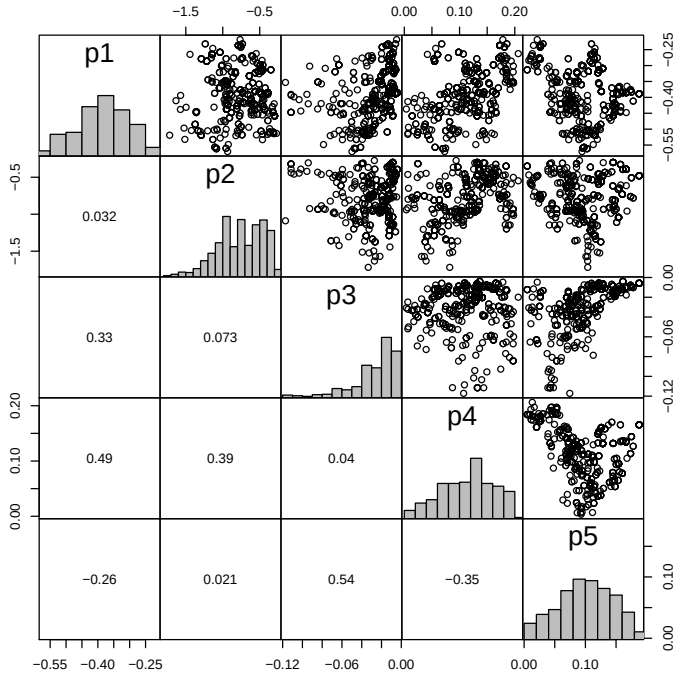


Figure 6: Posterior parameter distributions for the parameters of the model described by equation 2. p1= $k_1$, p2= $k_2$, p3= $k_4$, p4= $a_{21}$, p5= $a_{31}$. Numbers in the lower diagonal indicate the correlation coefficient between parameters.

```
> par(mar=c(5,5,4,1))
> plot(summary(sR),xlim=c(1950,2010),ylim=c(0,1000),xlab="Year",
+      ylab=expression(paste(Delta^14,"C ","(\u2030)")),main="")
> points(DataR14t,pch=20)
> lines(C14Atm_NH,col=4)
```

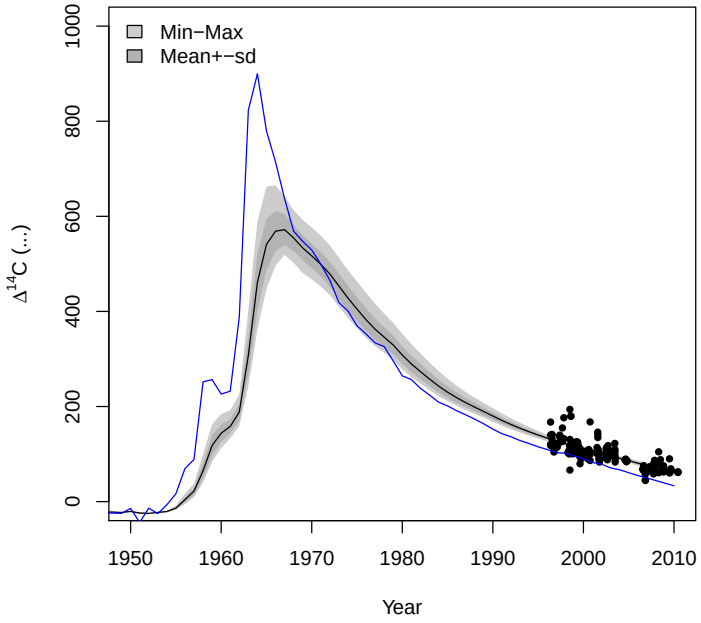


Figure 7: Predictions of respired radiocarbon values from the model of equation 2 versus observations. Model predictions include uncertainty range for the mean ± standard deviation, and the minimum-maximum range. Radiocarbon concentration in the atmosphere is depicted in blue.

## Acknowledgements

We would like to thank Nicole Gimbel for performing the laboratory incubation study. Susan E. Trumbore provided the radiocarbon data for Harvard Forest and gave important support and insights for the development of this project. Funding from the Max Planck Society.

## References

Schädel, C., Y. Luo, R. David Evans, S. Fei, and S. Schaeffer. Separating soil $CO_2$ efflux into C-pool-specific decay rates via inverse analysis of soil incubation data. *Oecologia* 171: 721–732.

Sierra, C.A., M. Müller, and S. E. Trumbore. 2012. Models of soil organic matter decomposition: the SoilR package, version 1.0. *Geosci. Model Dev.*, 5: 1045–1060.

Sierra, C.A., S. E. Trumbore, E. A. Davidson, S. D. Frey, K. E. Savage, and F. M. Hopkins. 2012. Predicting decadal trends and transient responses of radiocarbon storage and fluxes in a temperate forest soil. *Biogeosciences*, 9: 3013–3028.

Soetaert K. and T. Petzoldt. 2010. Inverse modelling, sensitivity and monte carlo analysis in R using package FME. *Journal of Statistical Software*, 33: 1–28.