DATA PREPROCESSING FOR PARAMETER ESTIMATION- AN APPLICATION TO A REACTIVE BIMOLECULAR TRANSPORT MODEL

[1]Cuch, Daniel A., [2]Rubio, Diana and [1]El Hasi, Claudio D.
[1]Instituto de Ciencias - Universidad Nacional de Gral Sarmiento, J.M. Gutierrez 1150, Los Polvorines, Buenos Aires, Argentina
[2]Centro de Matemática Aplicada, ECyT-Universidad Nacional de Gral San Martín. M. De Irigoyen 3100, 1650 San Martín, Buenos Aires, Argentina,
E-mail: drubio@unsam.edu.ar

Abstract: In this work we are concerned with the inverse problem of the estimation of modeling parameters for a reactive bimolecular transport based on experimental data that is non-uniformlydistributed along the interval where the process takes place.
We proposed a technic that helps to determine the intervals where most of the data should be considered in order to obtain a good estimation of the modeling parameters. For the purpose of reducing the cost of laboratory experiments, we propose to simulate data by means of a fitting function where are needed and are not available. We applied this strategy on the estimation of segregation and dispersion parameters for an advection-dispersion-reaction problem in a porous media. Each step is explained in detail and simulation results are shown and compared with previous ones.
Keywords: Transport, preprocessing, parameter estimation, segregation, mathematical modeling.

1. INTRODUCTION

When flow processes with multiple reactive species are studied, it is important to consider the nature of reactions and the role of fluctuations at small scale [1]. While the equations at Darcy´s scale are based on a continuum hypothesis, averaging over an elevated number of pores, the reaction dynamics is governed by poral scale processes [2]. In the smaller ranges the fluid velocity is never homogeneous in space, and the continuum hypothesis loses validity.In the flow of reactive solutes the continuum approach can produce erroneous results since reactants are considered homogenized in Darcy scale but they are not perfectly mixed at poral scale [3,4]. Here we use a model that changes the rate of reaction in the transport equation incorporating poral scale effects, or *segregation.*

Usually mathematical modeling of a physical process involves an inverse problem that consists in estimating one or more parameters of the model based on experimental data

(observations) [5]. Once the model is established, the parameters are usually estimated by minimizing the square errors between the experimental data and simulated values (mathematical model output). When data are obtained without previous analysis of where they should be taken in order to provide useful information for the parameter estimation, they are likely to provide information that is not enough for the correct modeling. In these cases, instead of repeating the experiment, a skilled data preprocessing may be appropriated.

In this paper we focus on numerical treatment in order to provide a simple and effective manner to model the problem of segregation. We consider a one-dimensional ADRE (advection-dispersion-reaction equation) model that includes two free parameters, one for the segregation term and the other for the dispersion. With the purpose to take into account informative data we propose to fit experimental data by a smooth function, so that we have data at any point.

Instead of using adapting steps procedures, we develop a simpler strategy that considers a shorter integration step in areas where the rate of change of the concentration of product is most significant. Numerical simulations were made and the segregation and dispersion parameters were estimated by analyzing the production profile and the mass of product.

2. MATHEMATICAL AND NUMERICALFORMULATION

We consider two reactive solutes A and B, having concentrations c_1 and c_2 respectively, in a transport flux where A displaces B at a fixed rate V and producing C at the interface, with concentration c_3. Assuming a stationary absorption process between the solid and liquid phases, and homogeneity of the reactive substances, the process may be considered as a bimolecular transport.

For simplicity, we assume that the reactive substances move along an enclosure of uniform section in which the dynamics along the direction of displacement is the only relevant one, so that it can be assumed that the transport process is one-dimensional. Although the velocity V, dispersion D, and the porosity ϕ can be space dependent, given the assumption of homogeneity they may be considered constant throughout the space-time region $\mathbf{\Omega}$ where the process takes place. The process can be described by a set of second order partial differential equations [6]:

$$\frac{\partial c_i(t,x)}{\partial t} + V\frac{\partial c_i(t,x)}{\partial x} - D^*\frac{\partial^2 c_i(t,x)}{\partial x^2} = R^*(t,x), (t,x) \in \Omega, \ i = 1,2,3 \tag{1}$$

where c_i are the solute concentrations, as mentioned above, for A, B and C, respectively, $D^* = D/\phi, R^* = R/\phi$ is the source term and $\Omega = [0,T] \times [0,L]$

The source term is modeled as a bimolecular reaction averaged over a large number of pores:

$$R = \Gamma_{eff} \overline{c_1 c_2} = \Gamma(1+s)\overline{c_1 c_2} \tag{2}$$

being s the segregation factor which is always negative since it reduces the reaction rate Γ [4]. Changes in concentration at poral scale are included in this correction as follows [7]:

$$s = \frac{\overline{c_1' c_2'}}{\overline{c_1 c_2}} \approx \frac{\alpha}{\emptyset} \frac{\overline{\nabla c_1} \overline{\nabla c_2}}{c_1 c_2} \tag{3}$$

Note that gradients, (in our case) have opposite signs since as one concentration increases the other decreases. Since s is negative one can conclude that the factor α must be positive.

Therefore, the mathematical model for the one-dimensional bimolecular reactive transport is

$$\frac{\partial c_k(t,x)}{\partial t} + V\frac{\partial c_k(t,x)}{\partial x} - D^*\frac{\partial^2 c_k(t,x)}{\partial x^2} = \Gamma\left(1+\frac{\alpha}{\emptyset}\frac{\overline{\nabla c_1}\overline{\nabla c_2}}{c_1 c_2}\right)\overline{c_1 c_2}, \quad k = 1,2,3, (t,x) \in \Omega \tag{4}$$

The equations are numerically solved using a finite difference scheme centered in space and forward in time, which guarantees a first order accuracy in time and second order in space. The discretization and numerical schemes and initial and boundary conditionsare the sameto the ones used in [8]. Here we include the resulting discretized equations.

Let $P = \{(t_i, x_j), i = 1, \dots, N_t, j = 1, \dots, N_x\}$ be a partition, uniform in x and in t, of the domain Ω, then the discretization of above equations takes theform

$$c_{k\,i+1,j} = c_{k\,i,j} - V\frac{\Delta t}{\Delta x}\left(c_{ki,j+1} - c_{ki,j-1}\right) + D^*\frac{\Delta t}{(\Delta x)^2}\left(c_{ki,j+1} - 2c_{ki,j} + c_{ki,j-1}\right) +$$

$$\Gamma\Delta t\left[1 + \frac{\alpha}{\phi}\frac{(c_{1i,j+1}-c_{1i,j-1})(c_{2i,j+1}-c_{2i,j-1})}{(2\,\Delta x)^2 c_{1i,j}c_{2i,j}}\right]c_{1i,j}c_{2i,j} \tag{5}$$

where $c_{ki,j} = c_k(t_i, x_j)$, $k = 1,2,3$, $i = 1, \dots, N_t, j = 1, \dots, N_x$, $\Delta t = t_{i+1} - t_i$ and $\Delta x = x_{j+1} - x_j$ (constant).

Note that the concentration $c_k(t_i, x_j)$ actually depends not only on time and position but also on the parameter values α, D, so we may write $c_k(t_i, x_j, \alpha, D)$.

We assume that there exists a pair (α_0, D_0) such that a set of M experimental data may be written as

$$d_n^{exp} = c_3(t_i, x_j, \alpha_0, D_0) + \varepsilon_n, n = 1, \dots M,$$

being c_3 the concentration of product and $\{\varepsilon_n\}$ independent realizations of a normal random variable with mean 0 and variance v.

The values of the parameters D and α are estimated from $\{d_n^{exp}, n = 1, \dots M\}$ by minimizing the square residuals, that is,

$$(\hat{\alpha}, \hat{D}) = \operatorname{argmin}_{D,\alpha} J(\alpha, D) = \sum_{n=1}^{M}\frac{\left(d_n^{exp}-C_3(\alpha,D)\right)^2}{M} \tag{6}$$

over all possible positive values for α and D, where $C_3(\alpha, D)$ is a vector of values of the simulated concentration c_3, obtained by solving the discretized equations (4), evaluated at the same time instants and places (t_i, x_j) at which the experimental data d_n^{exp} were taken.

The estimated pair $(\hat{\alpha}, \hat{D})$ is a realization of a random variable (α, D). It can be proved that under suitable hypothesis (α, D) has asymptotically normal distribution with mean (α_0, D_0) [9].

Note that $(\hat{\alpha}, \hat{D})$ depend on the number of data, M, the observed data d_n^{exp} (including the time and places where they were taken) and the initial values for (α, D) used to calculate $C_3(\alpha, D)$. Hence, besides the initial values (or *initial guess*) for the parameters, the accuracy of the estimation depend on the set of experimental observations, that is why is crucial to have informative data to yield to an accurate estimation of the parameters.

3. DATA PROCESSING

In practice, every time an experiment is performed a certain number M of experimental data become available, they might be sparse or not taken at the region of interest. Also experimental measurements contain uncertainties and an appropriated fitting function that approximate observed data may be considered without adding significant error. Afterwards, the resulting fitting curve could be used as *pseudo-experimental* data where is needed.

This is the idea for the strategy proposed in this work: use values of the fitted function as if they were experimental data. An analysis of the behavior of the particular process might help to choose the appropriated curve to be used for fitting.

The procedure proposed here, is based on the second derivative properties of a function. For a plane curve given by a twice differentiable function $y(x)$, the second derivative $y''(x)$ provide information about the acceleration or rate at which an object changes its velocity. In particular, in the case of the reactive-diffusive transport problem, for the curve that represents the amount of concentration of the product, the second derivative gives the rate of change in the velocity of production of concentration.

Assuming that a set of observation data $\{d_1^{exp}, \dots, d_M^{exp}\}$ is available to estimate a (vector) parameter θ, the procedure is a very simple one and it can be described in five steps.

1)   At a fixed time instant, fit the available data by a function $g(x)$. This can be done with the MATLAB function *fit* using one of the library models that are available.

2)   Define a uniform grid G in [0,L] and find the points $x_{mG} \in$ G where the absolute value of the second derivative of the fitting function $g(x)$ has local maxima, i.e.

$$x_{mG} = argmax\{|g''_{xx}(x)|, x \text{ in } G\}$$

3)   Define a new grid G^* by refining G around the points x_{mG}. For instance, consider $\Delta x^* = \frac{\Delta x}{10}$ as a step in the interval $[x_{mG} - \Delta x, x_{mG} + \Delta x]$. In this way, you generate a partition that is non-uniform in space.

4)   Calculate simulated data $c_j = c(x_j)$ at the new grid points $x_j \in G^*$, solving the numerical model for a particular value (*initial guess*) of the parameters θ. Note that c_j depends on that chosen value.

5)   Find the estimated value $\hat{\theta}$ of the (scalar or vector) parameter θ that minimizes the square error

$$\hat{\theta} = \text{argmin}_\theta J(\theta) = \sum_{j=1}^{N} \frac{\left(d_j^{FD} - c_j(\theta)\right)^2}{N} \qquad (7)$$

where $d_j^{FD} = g(x_j)$.

This procedure can be applied to a wide range of situations, even in cases with a large variability in the distribution of experimental data. In the profile data there might be areas of very rapid change, as in the bimolecular reaction process we are considering. In situations like this one, when using large integration steps some details of the profile can be lost, on the other hand, the computation becomes very slow if the step is too small. Hence, there is a trade of between accuracy and integration time, and this is an important issue to be considered when solving these problems numerically.

4. APPLICATIONS AND DISCUSSION

In order to use the preprocessing scheme described before, we first analyze the function to be used for fitting the experimental data.

For the problem of a bimolecular reactive transport problem discussed in section 2, the solution to the equation for the product has a bell shaped curve [3,6]. Hence, a Gaussian function is an appropriated one for fitting the experimental data for this problem. Based on this fact, the algorithm described in the previous section is implemented for the problem of the estimation of both, the segregation and the dispersion coefficient of the transport reactive – advection bimolecular problem where laboratory measurements are approximated by a Gaussian function.

We apply the proposed strategy to the experiments performed by [3,4]. The results are shown and discussed for each case.

Example 1.

As a first example we consider the process described in [3] and their experimental results. The concentration of product (c_3) at each point x of the column is reported at four fixed instants. The values of the characteristic constants of the experiment: reaction constant (Γ), tube length (L) and porosity (ϕ), are given. The results of the simulations are shown in Figure 1, wherein the concentration of product is observed for the incoming solution for a flow rate of V = 0.0121 cm/s. The red dotted lines correspond to the experimental data,the dashed lines corresponds to an analytical solution without considering the segregation [3] while the blue solid line is the result of the simulations obtained with the estimated parameters, after performing the fitting proposed.

It can be seen that the simulated curve fits well the experimental results and provide a good estimation of the concentration profile for c_3. Similar results are obtained for all the cases reported in [3].

We estimate both parameters, α and D, for each of the four time instants considered in [3].The results are shown in Table 1.

It can be observed that for different sets of data of the same experiment (taken at different times) might indicate a dependency on time for both parameters. Nonetheless, since they are very similar we could assume that they are constant in time for this problem (with the limitations imposed by the small number of cases) and equal to $\bar{\alpha} = 0,155$ and $\bar{D} = 0,0012$ (mean values). We observe that the estimated values $\hat{D}$ for the coefficient of dispersion are lower than the one determined in non reactive experiment ($D = 0.00175\ cm^2/s$) in [3]. However, the latter would give a narrower concentration profile [8].
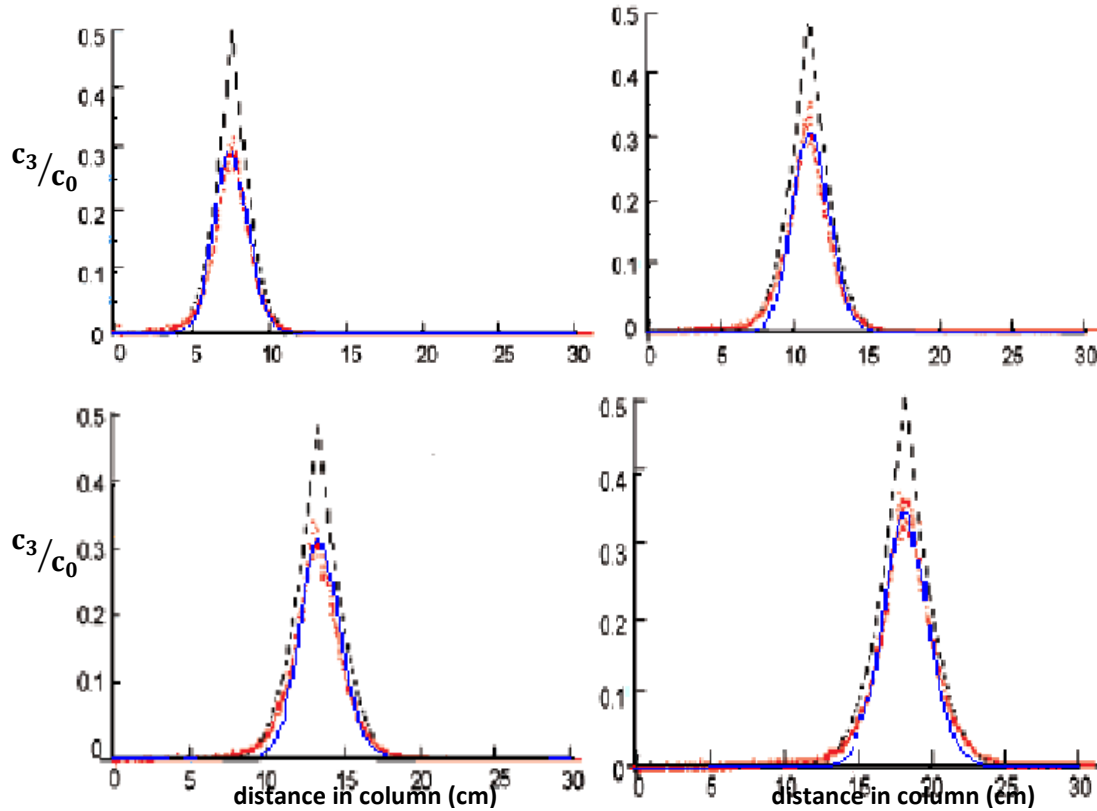
Figure No1. Product concentration with respect to the initial value along the column for time instants T = 619 s, 916 s, 1114 s, 1510 s. V = 0.0121 cm/s.
---- experimental data [3]   ----analitical solution [3]   ----numerical solution

T(s)	619	916	1114	1510
$\hat{\alpha}(cm^2)$	0.131	0.166	0.158	0.164
$\widehat{D}(cm^2/s)$	0.0013	0.0012	0.0011	0.0012

Table No1. $\hat{\alpha}$ and $\widehat{D}$ for T = 619 s, 916 s, 1114 s, 1510 s. V = 0.0121

Another indication of the reliability of the preprocessing method that we propose here is that we are able to estimate the total amount of mass produced during the transport process. We calculate the mass production considering the molecular weight of the concentration of product obtained in three ways: 1) experimental measurements, 2) simulated data using estimated values from experimental data and 3) simulated data using estimated values calculated after preprocessing data as explained in the previous section, where experimental measurements are fitted by a Gaussian function. The results are shown in Figure 2. It can be observed that a more accurate value results when the preprocessing method is applied.

Eventhough the resulting values are lower than the ones obtained from experimental measurements, they are very close to the expected ones.
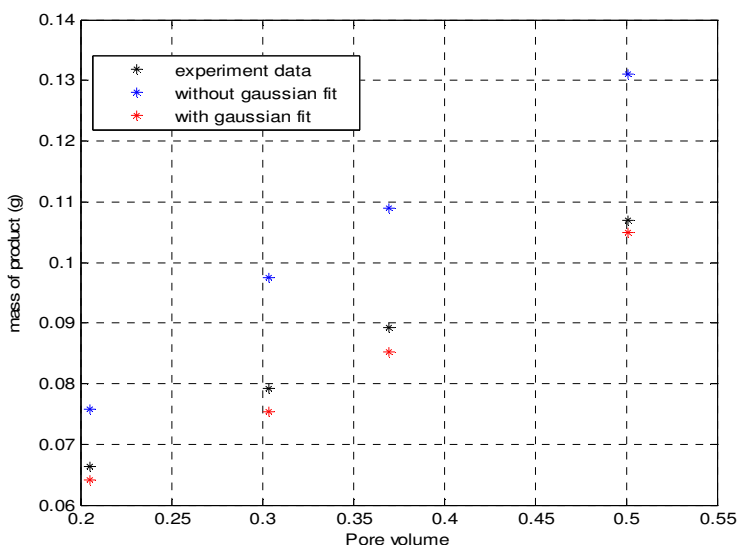

Figure No 2. Total mass of the product for T = 619 s, 916 s, 1114 s, 1510 s, for the experimental data [3] and different numerical simulations. V = 0.0121 cm/s.

Now, we consider other experimental settings reported in [3], with increasing values of the flux velocity (*V*). The estimated parameter values are shown in Table 2. It can be observed an increased value of the dispersion coefficient, which is an expected result, since at low flow rates the dispersion is proportional to the velocity ($D = \lambda V$, being λ the dispersivity).

V(cm/s)	0.0121	0.0832	0.67
$\hat{\alpha}(cm^2)$	0.155	0.196	0.23
$\widehat{D}(cm^2/s)$	0.0012	0.0116	0.13

Table N°2. $\hat{\alpha}$ and $\widehat{D}$ for different velocities: 0.0121 cm/s, 0.0832 cm/s and 0.67 cm/s.

We observe again that the estimated value $\widehat{D}$ corresponding to V = 0,0832 cm/s and V = 0,67 cm/s are lower than that determined in the non reactive experiment ($D = 0.0145\ cm^2/s$ and $D = 0.175\ cm^2/s$, respectively) in [3] being the differences about 25%. This would suggest that using D coefficient obtained in batch could lead to non-accurate results.

Example 2

Now we consider the experiment described in [4]. In this case the concentration of the product is measured at the end of the column (*x* fixed), as a function of time. The experiment

is performed at two different flow rates. The setup and the values of the parameters of interest are given in [4].

Figure 3 shows the results for the case of an initial concentration of 0.5 mM for both reactants A and B, and a flow rate V= 0.096cm/s. The squares in the figure correspond to the experimental data, the black line corresponds to the results obtained withou tconsidering the segregation effects and the blue line corresponds to the simulated result obtained using the preprocessing strategy introduced in section 3.
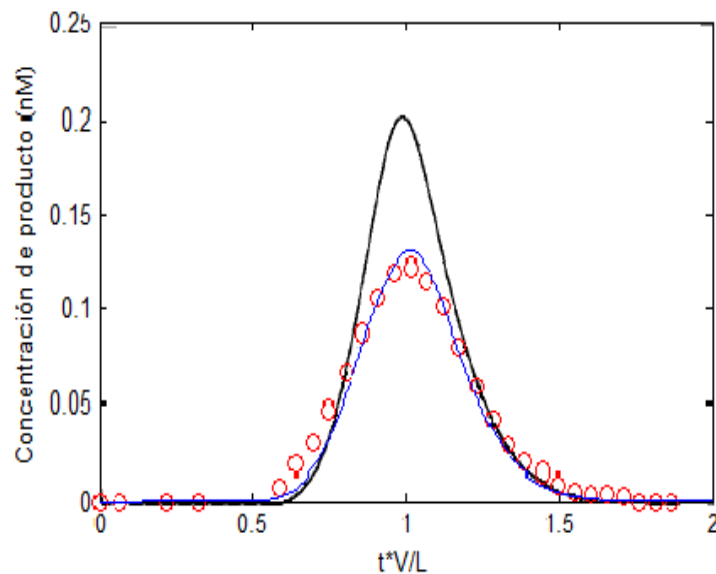

Figure No 3. Product concentration at the end of the column as a function of the time. Initial concentration for reactants: 0.5 nM, flow velocity: 0.096 cm/s.
___ numerical solution without Gaussian fit and parameters adjust [4]
ooo experimental data [4]   ___ our numerical solution

Figure 4 shows the results for an initial concentration of 0.25 mM for reactants A and B anda flow rate of 0.07 cm/s.

In both cases the numerical simulations fit very well the experimental data. For the simulated results we use the estimated parameter values $\hat{\alpha}$ and $\hat{D}$ for the segregation coefficient (α) and for the dispersion coefficient (D) shown in Table 3.

In this example, the estimated value $\hat{D}$ for the coefficient of dispersion is a little higher than the one determined in the non reactive experiment ($D = 0.023\ cm^2/s$ and $D = 0.032\ cm^2/s$, respectively) in [4], the differences are about 7 %. Once more we note that the dispersion coefficient increases with speed, as it was expected.
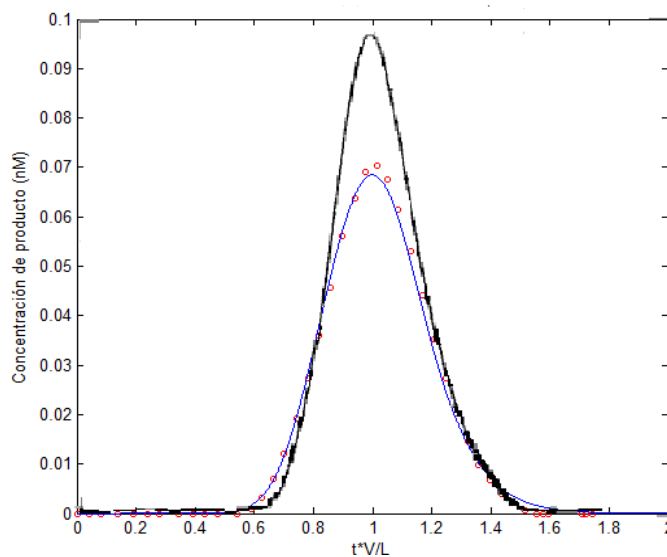
Figure No 4. Product concentration at the end of the column as a function of the time. Initial concentration for reactants: 0.25 nM, flow velocity: 0.07 cm/s.
___ numerical solution without Gaussian fit and parameters adjust [4]
ooo experimental data [4]   ___ our numerical solution

V	0.07 cm/s	0.096 cm/s
$\hat{\alpha}(cm^2)$	2	1.81
$\hat{D}(cm^2/s)$	0.025	0.034

Tabla No 3. $\hat{\alpha}$ and $\hat{D}$ for an initial concentrations of 0.25 nM and 0.5 mM for the reactants A and B.

In both experiments considered in this work, we can observe the dependency of the dispersion coefficient on speed. The values for D obtained for the [3] case are smaller than the ones of the experiments with nonreactive flow, reported in the original paper. For the ADRE model it is assumedthat the transport process obeys the Fick's Law ($j = -D \, \nabla c$). As the reaction time is several orders of magnitude shorter than that of the advection and dispersion in the region of the reaction front, reactive substances are quickly consumed resulting in very large concentration gradients,this could explain the difference between the estimated values $\hat{D}$ and the one calculated in batch. On the other hand, in the experience of [4], all characteristic times are of the same order of magnitude while the gradients are lower, which could explain why the value for $\hat{D}$ is similar to the one determined in the nonreactive experiment.

These observations regarding the gradients of the reactants are in accordance with equation (3). The gradients, and its product, are greater in [3] than in [4], which might explain why the estimated parameter $\hat{\alpha}$ is smaller in the first example.

4. CONCLUSIONS

It is often difficult to have experimental data in the whole region of interest. They might be either scarce or not enough in the areas where the process to study present great variations or contain more information about the process. By means of a simple method that generates a smooth curve that approximates the experimental data, we can get "simulated data" that allow us to reproduce the experiments and, for instance, build an appropriated space discretization reducing integration step only in areas of interest. This methodology is simple, fast and the results presented here indicates its efficiency.

Furthermore, the use of macroscopic models (continuous), although it has its limitations, it may be useful to analyze, quickly and easily, a number of phenomena. Herewe use the ADRE approach to analyze experimental results, showing some discrepancies between the model and the experiments that deserve further analysis. However the accurated estimation of modeling parameters using pseudo-experimental data obtained by the methodology presented here,appears as a useful tool for modeling and forecasting in different situations.

REFERENCES

[1] Edery, Y., Guadagnini, A, Scher, H and Berkowitz, B. 2013. *Advances in Water Resources*, **51**, 86-103, and references therein.

[2] Kapoor, V., Gelhar, L.W. and Miralles-Wilhelm, F., 1997, Bimolecular second-order reactions in spatially varying flows: Segregation induced scale-dependent transformation rates. *Water Res. Research.* 33 (4), 527-536.

[3] Gramling, C., Harvey, C. and Meigs, L., 2002. Reactive Transport in Porous Media: A Comparison of Model Prediction with Laboratory Visualization. *Environmental Science & Technology.* 36 (11), 2508-2514.

[4] Raje, D. S. and Kapoor, V., 2000, Experimental Study of bimolecular Reaction Kinetics in Porous Media. *Environ. Sci. Technol*. 34, 1234-1239.

[5] Tarantola, A, 2005.Inverse problem theory and methods for model parameter estimation. *SIAM*.

[6] Bear, J., 1988. Dynamics of fluids in Porous Media. Dover Publications, Inc. New York.

[7] Meile C. And Tuncay K., 2006. Scale dependence of reaction rates in porous media. Adv. in *Water Resources.* 29(1), 62-71.

[8] Rubio, A.D., Zalts, A. and El Hasi, C.D., 2008*,* Numerical solution of the advection reaction diffusion equation at different scales. *Environmental Modelling & Software.* 23, 90-95.

[9] Seber, G.A.F. and Wild, C.J. 2003. Nonlinear regression, Wiley Intersciences, NJ.