Biomatemática ***26*** *(2016), 97–120* ***ISSN 1679-365X***
Uma Publicação do Grupo de Biomatemática IMECC – UNICAMP

Relevância do tratamento de separação de dados para a estimativa do coeficiente de difusão em problemas ecológicos

Camile F.D. Kunz[1], Juliana M.R. Souza[2],
João F.C.A. Meyer[3],
IMECC, UNICAMP, 13.083-970, Campinas/SP.

Resumo. O objetivo realizado por esse trabalho foi avaliar o impacto do tratamento de separação de dados para a estimativa do coeficiente de difusão/dispersão para problemas ecológicos e epidemiológicos. Em um trabalho anterior, foi desenvolvido e validado um método para a estimativa do coeficiente de difusão, considerando que havia apenas um foco de eventos relevantes. A validação de tal método lançou mão de técnicas de simulação de dados de caráter difusivo. No presente trabalho, como exemplo, é estudado o caso em que os eventos possuem dois focos de eventos, geograficamente separados. Dois experimentos distintos, no tocante à geração de dados, foram realizados e detalhadamente estudados, utilizando o algoritmo de clusterização K-means. Ao final, é feita uma aplicação prática, a gripe aviária na Nigéria, utilizando dois algoritmos de clusterização distintos para lidar com dados em latitude e longitude.

Palavras-chave*: Coeficiente de Difusão, Problemas Ecológicos, Estimativa de Parâmetros, Data clustering*

1. Introdução

Em (Souza, 2014), desenvolveu-se um método para a estimativa do coeficiente de difusão para problemas ecológicos. A especificidade da classe de

[1]camileknz@gmail.com
[2]jumarta@gmail.com
[3]joni@ime.unicamp.br

problemas à qual o método se destina se deve às particularidades dos dados de tal área.

As hipóteses a serem tomadas, para que o modelo seja coerente com a realidade, são distintas dos demais problemas de difusão, de outras áreas. Em modelos difusivos nos quais as partículas são moléculas, ou alguma substância química, supõe-se que o processo difusivo decorre da aleatoriedade do movimento. Nesse caso, tem-se como hipótese que não há interação entre as partículas; assim, o movimento de uma única partícula não depende do movimento de outras.

Em problemas ecológicos, muitas vezes, o processo difusivo pode ser decorrente da interação entre partículas. Nesse caso, trata-se por partículas o objeto de estudo, que pode vir a ser indivíduos, ou o risco de infecção por um vírus, por exemplo.

Outra particularidade da área de interesse se encontra na maneira de se coletar informações. Quando o objeto de estudo é de nível molecular é possível observar o processo de forma controlada. Através de equipamentos, cria-se um sistema isolado do mundo e o objeto de estudo é manipulável. No entanto, para problemas ecológicos, obter informações pode ser uma tarefa árdua.

É laborioso delimitar o sistema a qual o objeto de estudo pertence e, desta maneira, conseguir coletar informações que sejam confiáveis. E, também, ao resgistrar ocorrências de interesse, nada impede que hajam informações perdidas ou ocultas, impossibilitando o conhecimento completo acerca do processo difusivo.

Em (Souza, 2014), a geração de dados, com um coeficiente de difusão supostamente conhecido, foi tão relevante quanto o próprio desenvolvimento do método. Desta forma, decorrente da dificuldade de obtenção de dados confiáveis originados de processos difusivos em problemas ecológicos, Souza (Souza, 2014), optou por simular tais dados para a devida validação do método. Tal abordagem considerou a existência de apenas um foco de eventos de interesse.

No entanto, para o caso de dados reais, pode ter ocorrido mais de um foco de eventos. A ocorrência ou não de mais de um foco pode ser avaliada visualmente. E, caso a distribuição dos eventos não esteja claramente separada para se observar o número de focos, é possível estimar o coeficiente de difusão para diversas clusterizações. A pertença atribuída aos dados não é única e independente, necessita da interpretação do espectador que estuda o processo.

Desta forma, o objetivo do presente trabalho é avaliar o impacto de se

considerar de multiplos focos na recuperação do coeficiente de difusão original por trás de cada processo. Além disso, avaliar como o tratamento dos dados via clusterização de acordo com sua posição espacial pode melhorar tais estimativas via método desenvolvido por (Souza, 2014). Também é estudada uma aplicação prática do método para a estimativa do coeficiente de difusão do H5N1 na Nigéria.

2. Metodologia

Na tese de Doutorado de Souza (Souza, 2014), foi desenvolvido e validado um método para a recuperação do coeficiente de difusão para problemas ecológicos para o caso em que havia apenas um foco de eventos. Tal método pode ser aplicado a conjuntos de dados reais de caráter puramente difusivo, isto é, casos em que o espalhamento devido à velocidade de transporte seja inexistente ou desprezível.

Com o intuito de validar tal técnica, foi preciso testá-la em um conjunto pertinente de dados . Para isso, em (Souza, 2014), foram construídas ferramentas computacionais que possibilitassem a geração de tais dados, para a validação do método de obtenção de α, o coeficiente de difusão.

No presente trabalho, colocam-se as seguintes questões: e se os dados em mãos foram originados de dois ou mais focos? É possível aplicar o método desenvolvido em (Souza, 2014) a esses dados tais como estão? Se não, qual o tratamento adequado para obter resultados fidedignos nesse caso? Ou seja, almeja-se aqui estudar a relevância do tratamento de separação de dados previamente à recuperação do coeficiente de difusão via técnica em (Souza, 2014).

A seguir serão descritos o método desenvolvido e a técnica utilizada para sua validação. Também, serão descritos os algoritmos de separação de dados empregados nesse trabalho.

2.1. Revisão metodológica

Nessa seção, será explicitada a construção da técnica desenvolvida por (Souza, 2014) para a geração de dados de problemas difusivos. Para um coeficiente de difusão α, obtém-se a solução numérica do Problema de Valor Inicial e de Contorno de interesse através da construção de base de Elementos Finitos e discretização temporal via Método de Crank-Nicolson.

E, a partir dessa solução, constrói-se uma função de densidade de probabilidade conveniente, que é utilizada para a geração dos dados desejados. Tais dados são utilizados para a recuperação do coeficiente de difusão através da equação de ajuste:

$$\left(\sqrt{T_1} + \sqrt{t}\right) X_1\, x - \left(\sqrt{t} - \sqrt{T_1}\right) X_1^2 = 4\, T_1\, \sqrt{t}\, \alpha \tag{2.1}$$

Nessa equação, T_1 e X_1 são as coordenadas mais próximas do foco eventos. O termo evento é utilizado para denotar um par (t, x), seja ele obtido através de simulações de dados, ou seja ele um registro de ocorrência no mundo real. Desta forma, o foco de eventos será o primeiro evento evento registrado.

Tendo em vista o objetivo de validação do método de obtenção do coeficiente de difusão, em (Souza, 2014), utilizou-se a geração de dados proveniente de um problema difusivo. Para tal geração, foi empregado o Método da Transformada Inversa (Ross, 2006):

> Seja U uma variável aleatória contínua no intervalo $[0, 1]$. Então, para qualquer função de distribuição contínua F, a variável aleatória X definida por
>
> $$X = F^{-1}(U)$$
>
> tem distribuição acumulada F.

Desta forma, através de uma distribuição uniforme $U \sim [0, 1]$ e de uma distribuição acumulada, é possível obter uma variável aleatória X relativa à distribuição F.

Uma função f é densidade de probabilidade de uma variável aleatória contínua se satisfaz as condições:

i) $f(x) \geq 0$ para todo $x \in \mathbb{R}$

ii) $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1$

A *função de distribuição acumulada* $F(x)$ representa a probabilidade de que o valor da variável aleatória seja menor ou igual a x, ou ainda,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

A função F satisfaz as seguintes propriedades:

- F é não decrescente
- F é contínua
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to +\infty} F(x) = 1$

Com o intuito de obter uma distribuição acumulada condizente ao problema difusivo, em (Souza, 2014) foi proposto o seguinte Problema de Valor Inicial e de Contorno (PVIC):

$$\begin{cases} \dfrac{\partial c}{\partial t} = \alpha \dfrac{\partial^2 c}{\partial x^2}, \quad \forall\, x \in (x_0, x_f)\, e\, t \in (t_0, t_t] \\ \dfrac{\partial c}{\partial x}\Big|_{x \in \{x_0\} \cup \{x_f\}} = \dfrac{\beta}{\alpha}\, c(x), \quad \forall\, t \in (t_0, t_t] \\ c(x, t_0) = f(x), \quad \forall\, x \in [x_0, x_f] \end{cases} \tag{2.2}$$

Em (Souza, 2014) utilizou-se o Método de Elementos Finitos combinado ao Método de Galerkin, para o tratamento da componente espacial, com ordem de convergência $O(\Delta x^2)$. Para a a discretização temporal, utilizou-se o método de Crank-Nicolson cuja ordem de convergência é $O(\Delta t^2)$. Obtém-se C_i^n para cada passo de tempo n, para $n = 1, \ldots, Nt$ e $i = 1, \ldots, Nx$.

Observa-se que $\min_i\{C_i^n\} \geq 0$, pois concentrações negativas não fazem sentido no modelo do problema ecológico. Assim, a condição *i)* de uma função de densidade de probabilidade é satisfeita. Agora, é preciso normalizar C^n para satisfazer à condição *ii)*. Mantendo a notação C^n para a normalizada,

$$C^n = \frac{C^n}{\int_{x_1}^{x_{\bar{N}}} C^n\, dx} \tag{2.3}$$

A função de densidade de probabilidade C^n obtida é linear por partes.

Para obter a densidade, é feita interpolação linear entre cada par de pontos (x_{i+1}, C_{i+1}^n) e (x_{i+2}, C_{i+2}^n). Em seguida, integra-se a função interpolante

para $x \in (x_{i+1}, x_{i+2})$ e o resultado é a acumulada $\bar{C}^n(X \leq x)$.

$$\begin{aligned}\bar{C}^n(x) = \frac{C_1^n\, h}{2} + h \sum_{j=2}^{i} \frac{C_j^n\, h}{2} + \frac{C_{i+1}^n\, h}{2} + \\ (x - x_{i+1})\Big[C_{i+1}^n + \frac{(C_{i+2}^n - C_{i+1}^n)}{2\, h}\,(x - x_{i+1})\Big]\end{aligned} \tag{2.4}$$

A função (2.4) é a distribuição acumulada utilizada no Método da Tranformada Inversa para a geração de dados.

O coeficiente de difusão recuperado, denotado $\bar{\alpha}$, é obtido através de (2.1) e depende diretamente das coordenadas do primeiro evento ocorrido. Desta forma, considera-se o primeiro evento ocorrido t_1 com localização x_1 como foco e determina-se os outros eventos com relação a este.

Tendo isso em vista, é realizado o seguinte tratamento para os dados gerados, em que N_e é o número de eventos totais obtidos através da simulação de dados:

$$(T_{i-1}, X_{i-1}) = (t_i, x_i) - (t_1, x_1), \qquad i = 2, \ldots, N_e \tag{2.5}$$

Outra questão importante, é que no caso de dados reais, não é possível prever o instante em que ocorrem eventos. Para simular essa incerteza de ocorrência entre eventos, define-se o parâmetro denotado por DT, que será o intervalo máximo permitido entre ocorrência de sucessivos.

O parâmetro DT representa a capacidade de vigília acerca do registro de ocorrências do objeto de estudo.

Empregado-se o parâmetro DT, a geração dos eventos é representada da seguinte maneira:

$$\begin{cases} \hat{t}_1 \in [1, DT] \\ \hat{t}_2 \in [\hat{t}_1 + 1, \hat{t}_1 + 1 + DT] \\ \vdots \\ \hat{t}_{N_e} \in [\hat{t}_{N_e} + 1, \hat{t}_{N_e} + 1 + DT] \end{cases} \tag{2.6}$$

2.2. Tratamento de separação de dados

Em (Souza, 2014), as ferramentas foram desenvolvidas supondo que os eventos no espaço-tempo foram originados a partir de um foco. No entanto, nos problemas da natureza, às vezes há mais de um foco de eventos e aplicar

o método em tais casos pode resultar em valores que não condizem com a realidade.

Na prática, através do gráfico de dispersão, é possível estimar aproximadamente o número de focos. Desta forma, o objetivo doravante será analisar o impacto do tratamento de separação de dados na recuperação do coeficiente de difusão, para o caso em que sabidamente houveram dois focos.

No contexto de agrupamento de dados, define-se o conceito de *Data Clustering* de acordo com Gan et al. (1979):

> Data Clustering (ou apenas clustering) é um método para criar grupos de objetos, ou clusters, tal que os objetos que pertençam ao mesmo grupo são muito similares e os objetos em grupos distintos muito diferentes.

Um conjunto de dados com m objetos, cada qual descritos por p atributos, é denotado por $D = \{x_1, x_2, \ldots, x_m\}$ em que $x_i = (x_{i\,1}, x_{i\,2}, \ldots, x_{i\,p})^T$ é o vetor que representa o *i-ésimo* objeto.

Mensurar a similaridade ou dissimilaridade entre objetos e então classificá-los, requer a definição de uma métrica. Para o caso dos dados trabalhados, em que o objetivo é separá-los no espaço unidimensional, define-se a função de distância:

$$d(i,j) = |x_i - x_j| \tag{2.7}$$

Em (2.7), $d(i,j)$ satisfaz as propriedades de uma métrica:

i) $d(x_i, x_j) \geq 0 \quad \text{e} \quad d(x_i, x_j) = 0 \iff x_i = x_j.$

ii) $d(x_i, x_j) = d(x_j, x_i).$

iii) $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j).$

Para a avaliação do impacto da separação de dados, escolheu-se um algoritmo de clusterização baseado em centros: o K-means, que descrito primeiramente por Macqueen (1967). Por ser de simples implementação e rápida convergência, é aplicado em diversos tipos de dados.

Tal método de clusterização requer o número de partições desejadas como entrada. No caso dos experimentos computacionais desse trabalho, tal característica é vantajosa, pois foi estudado o caso em que haviam dois focos de concentração.

De acordo com (Jain, 2010), o algoritmo K-means pode gerar partições vazias, o que pode indicar o caso em que o número de partições escolhido foi superestimado para o conjunto de dados. Outra característica, é que a convergência da partição depende muito da inicialização do centro dos clusters.

O algoritmo K-means inicializa-se com k centros aleatórios dos *clusters*. Em seguida, calculam-se as distâncias dos n objetos aos k centros. O objeto é atribuído ao *cluster* cuja distância é mínima. Então, os centros são atualizados e realiza-se o procedimento de atribuição novamente. O critério de parada do algoritmo é não haver mais atualização dos centros.

2.3. Abordagem da aplicação prática

Como será exposto adiante, foi realizada uma aplicação prática, o caso da gripe aviária na Nigéria. Tendo em vista que os dados de entrada se encontram em formato latidude/longitude, não é possível aplicar o algoritmo K-means com a função de distâncias Euclidianas (2.7).

Para definir uma métrica adequada, utiliza-se a função que calcula a distância entre pontos que se encontram em uma esfera. Utilizando a notação ϕ_1 e ϕ_2 para as latitudes dos pontos 1 e 2, λ_1 e λ_2 para suas longitudes, R_T para o raio da terra, a distância d entre os pontos será dada pela equação (2.8), de acordo com (CodeCodex, 2015).

Observa-se que para aplicar essa equação é preciso que as coordenadas estejam em radianos.

$$d \;=\; 2\,R_T \operatorname{arcsen}\left(\sqrt{\operatorname{sen}^2\left(\frac{\phi_1-\phi_2}{2}\right) \;+\; \cos(\phi_1)\cos(\phi_2)\operatorname{sen}^2\left(\frac{\lambda_1-\lambda_2}{2}\right)}\right) \tag{2.8}$$

O algoritmo K-means adaptado para dados em latitude/longitude inicializa os centros dos *clusters* com coordenadas aleatórias. Em seguida, é calculada a distância de cada ponto até o centro através de (2.7) e, assim como o K-means original, atribui-se o ponto ao *cluster* em que essa distância é mínima.

Em seguida, os novos centros são calculados e o processo se repete até que não haja atualização dos *clusters*.

Também foi utilizado o algortimo de clusterização utilizado por Souza em sua Dissertação de Mestrado (Souza, 2010). Tal algortimo foi desenvolvido para lidar com dados de latitude e longitude e também utiliza como métrica a função de distância (2.7).

Nesse algoritmo, inicialmente cada objeto pertence a um *cluster*. Calcula-se a distância entre todos os centros e os *clusters* cuja distância é mínima se unem. Tal procedimento é realizado até que o número de *clusters* seja k.

3. Ensaios numéricos

Em (Souza, 2014) foi validado um método para recuperação do coeficiente de difusão para o caso em que, supostamente, apenas um foco originou o deslocamento de concentração. Aqui, a pretensão é estudar o caso em que podem ter ocorrido dois focos de eventos.

Desta forma, o intuito destes experimentos é verificar a relevância do tratamento de separação de dados prévia à recuperação do coeficiente de difusão.

Assim, almeja-se verificar uma técnica de separação de dados , isto é, se a clusterização é eficiente para a recuperação do coeficiente de difusão.

Doravante, ao gerar diferentes conjuntos de dados e então aplicar o algoritmo de separação, espera-se que tal tratamento obtenha uma partição satisfatória com relação à partição original.

Posto que se sabe à qual conjunto o ponto originalmente pertence, é possível avaliar a eficiência do algoritmo de separação de dados.

Outro aspecto a ser analisado, ou critério de fidedignidade, tão importante quanto a eficiência de clusterização, é a recuperação do coeficiente de difusão para os dados separados. Tais coeficientes, obtidos após tratamento de separação, serão comparados com o obtido considerando dados com ocorrência de apenas um foco para verificar a relevância desse tratamento de dados.

Tendo em vista tais objetivos, define-se as seguintes notações para a realização dos experimentos:

- α: coeficiente de difusão utilizado na obtenção da solução numérica com os focos em locais distintos.
- α_1 e α_2: coeficientes de difusão estimados considerando a amostra isolada em um domínio com apenas um foco de eventos.
- $\bar{\alpha}_1$ e $\bar{\alpha}_2$: coeficientes estimados para cada partição obtida pela clusterização.
- $\hat{\alpha}$: coeficiente de difusão estimado para os dados combinados sem o particionamento.

Tem-se como hipótese de que a qualidade do algoritmo de clusterização é proporcional à distância $d = |\bar{x}_1 - \bar{x}_2|$, ou seja, quanto maior a distância entre os focos, maior a eficiência da clusterização.

Tendo isso em vista, as localizações dos focos foram escolhidas apenas em função da distância entre elas.

Assim, buscou-se um referencial em que a localização exata de $\bar{x}_1$ e $\bar{x}_2$ não fosse relevante e, sim, apenas d. Considerando o domínio espacial $[0, x_f]$, foi convencionado que a escolha da localização dos focos, $\bar{x}_1$ e $\bar{x}_2$, sendo $\bar{x}_1 < \bar{x}_2$, obedeceria à seguinte relação:

$$\bar{x}_1 = \frac{x_f}{2} \quad , \quad \bar{x}_2 = \frac{x_f}{2} + d \tag{3.9}$$

Desta forma, para dado coeficiente de difusão α obtém-se solução numérica, com passos de tempo sempre $Nt = 512$, $t_0 = 0$ e $t_f = 1$, domínio espacial $[0, x_f]$, para focos localizadas em $\bar{x}_1$ e $\bar{x}_2$.

Assim, a partir das soluções numéricas obtidas, utiliza-se o método de simulação de dados difusivos e, para dado conjunto de DT's, obtém-se um cenário para cada um dos focos.

Em seguida, os coeficientes de difusão α_1 e α_2 são estimados para cada um dos focos. Observe: até aqui, as concentrações são analisadas como ocorrências disjuntas, sem nenhuma relação entre si. Aqueles valores obtidos para α_1 e α_2 serão tidos como referência para os obtidos após o tratamento.

Adiante, os dados são misturados e o algoritmo de *cluster* é aplicado e classifica os dados de acordo com a coordenada espacial. É então que se recupera os coeficientes de difusão $\bar{\alpha}_1$ e $\bar{\alpha}_2$ para cada conjunto obtido pela classificação. É usada a convenção de que $\bar{\alpha}_1$ é o coeficiente para os dados do *cluster* de menor centro e $\bar{\alpha}_2$ de maior. Em seguida, é estimado $\hat{\alpha}$, o coeficiente para os dados com a ocorrência de apenas um foco, supostamente.

Também, com o intuito de medir a eficiência da separação, foi calculado o percentual médio de acertos do algoritmo para cada caso. Outro fator importante é a variação da distância d entre os focos, por isso foram analisados os casos em que $(2,5 \times 10^{-2})\, x_f \leq d \leq (9 \times 10^{-1})\, x_f$.

Para obter os cenários das próximas tabelas foi utilizado o software Matlab com vetor de DT's fixo tal que DT's$= [1 : 5 : 50]$. O algoritmo de *cluster* foi modificado de um programa original desenvolvido por Cao (2008).

Foram realizados dois experimentos computacionais distintos no tocante à geração de dados, e serão expostos a seguir.

3.1. Primeiro experimento e resultados

A figura (1) esquematiza os domínios espaciais utilizados na obtenção da solução numérica e para a representação dos dados mesmo domínio.

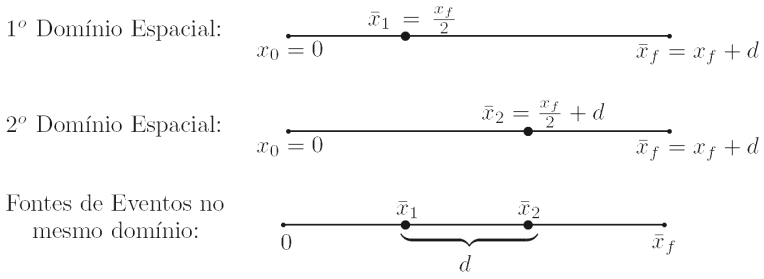

Figura 1: Esquema da localização dos focos para a obtenção da solução numérica e para a representação dos dados em mesmo domínio para o Primeiro Experimento.

No presente experimento, em todos os casos, foram tomados DT's fixos. Almejou-se analisar o impacto da inicialiazação aleatória ou fixa do algoritmo de clusterização na obtenção de $\bar{\alpha}_1$ e $\bar{\alpha}_2$.

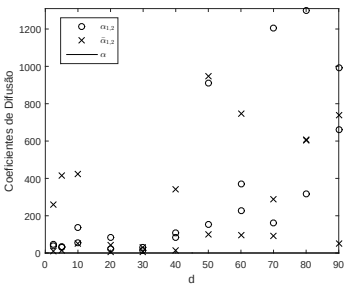

Figura 2: $\alpha = 1$. Distâncias versus $\alpha_{1,2}$ e $\bar{\alpha}_{1,2}$.

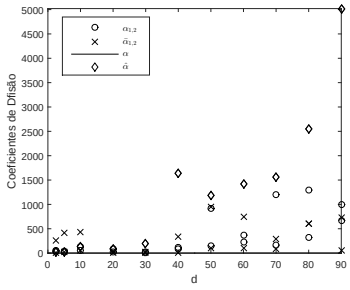

Figura 3: $\alpha = 1$. Distâncias versus $\alpha_{1,2}$, $\bar{\alpha}_{1,2}$ e $\hat{\alpha}$.

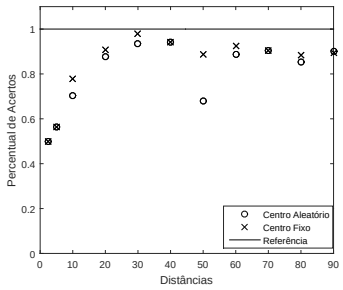

Figura 4: $\alpha = 1$. Acertos da clusterização para centro aleatório e fixo.

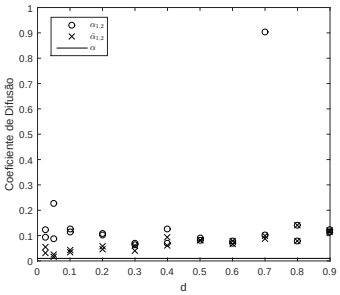

Figura 5: $\alpha = 10^{-2}$. Distâncias versus $\alpha_{1,2}$ e $\bar{\alpha}_{1,2}$.

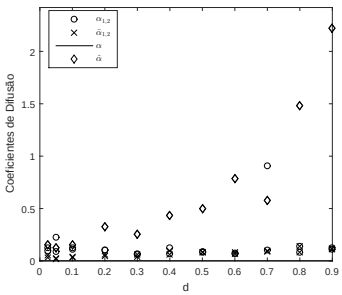

Figura 6: $\alpha = 10^{-2}$. Distâncias versus $\alpha_{1,2}$, $\bar{\alpha}_{1,2}$ e $\hat{\alpha}$.

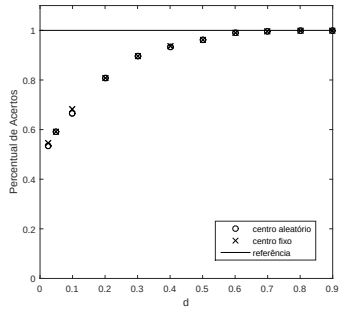

Figura 7: $\alpha = 10^{-2}$. Acertos da clusterização para centro aleatório e fixo.

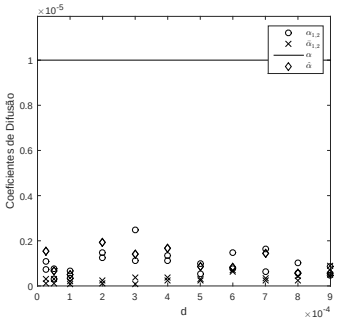

Figura 8: $\alpha = 10^{-5}$. Distâncias versus $\alpha_{1,2}$ e $\bar{\alpha}_{1,2}$.

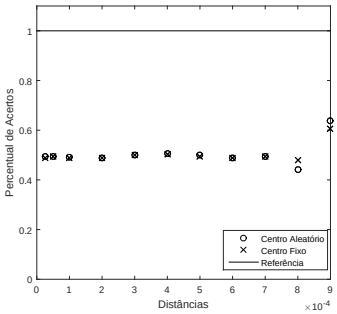

Figura 9: $\alpha = 10^{-5}$. Acertos da clusterização para centro aleatório e fixo.

3.2. Segundo experimento e resultados

Tendo em vista que o primeiro experimento foi um ponto de partida, neste segundo experimento almejou-se corrigir os problemas apresentados pelo anterior. A obtenção da solução numérica no domínio aumentado $[0, x_f + d]$ não retratou os dados de forma adequada, pois o coeficiente de difusão foi superestimado para $\alpha = 1$.

A figura (10) esquematiza os domínios espaciais utilizados na obtenção da solução numérica e para a representação dos dados mesmo domínio.

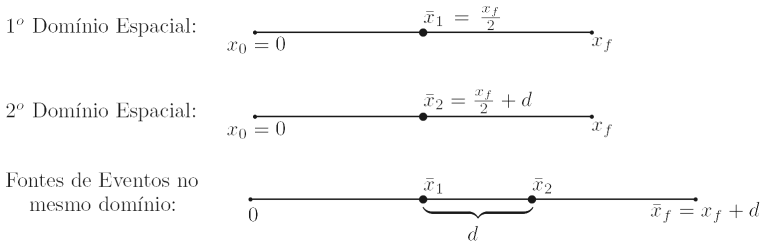

Figura 10: Esquema da localização dos focos para a obtenção da solução numérica e para a representação dos dados em mesmo domínio para o Segundo Experimento.

Outro aspecto analisado neste experimento foi o contraste entre considerar cenários com *DT*'s fixos ou aleatórios, assim, foram ensaiados ambos os casos. Como no experimento anterior já havia sido analisada a dependência

da inicialização do centro do *cluster*, aqui foi considerado apenas o caso com inicialização fixa e igual à localização dos focos.

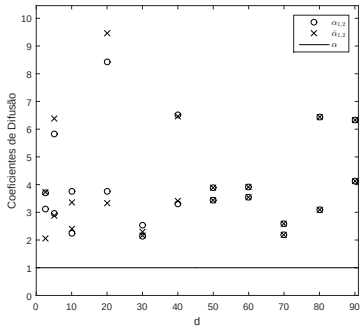

Figura 11: $\alpha = 1$. Distâncias versus $\alpha_{1,2}$ e $\bar{\alpha}_{1,2}$.

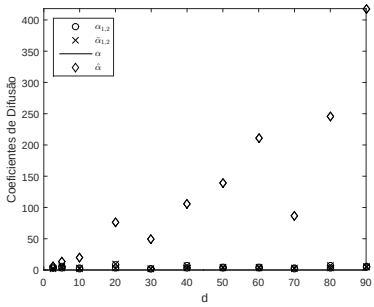

Figura 12: $\alpha = 1$ Distâncias versus $\alpha_{1,2}$, $\bar{\alpha}_{1,2}$ e $\hat{\alpha}$.

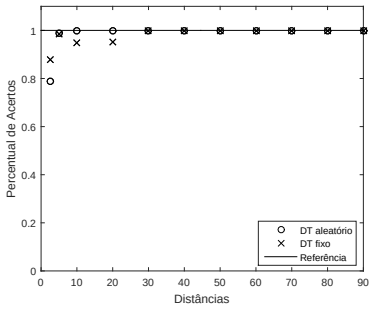

Figura 13: $\alpha = 1$. Acertos da clusterização para centro aleatório e fixo.

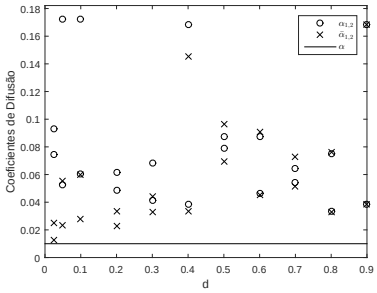

Figura 14: $\alpha = 10^{-2}$. Distâncias versus $\alpha_{1,2}$ e $\bar{\alpha}_{1,2}$.

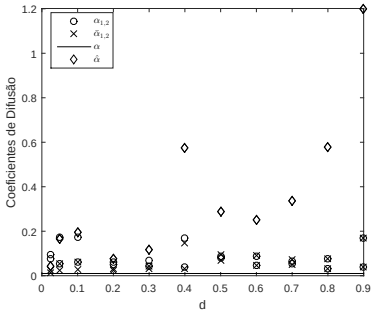

Figura 15: $\alpha = 10^{-2}$. Distâncias versus $\alpha_{1,2}$, $\bar{\alpha}_{1,2}$ e $\hat{\alpha}$.

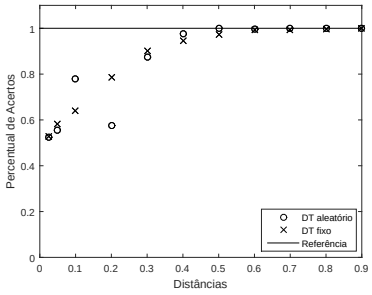

Figura 16: $\alpha = 10^{-2}$. Acertos da clusterização para centro aleatório e fixo.

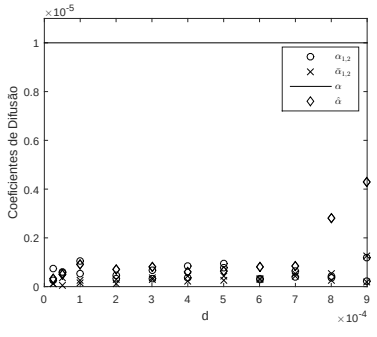

Figura 17: $\alpha = 10^{-5}$. Distâncias versus $\alpha_{1,2}$ e $\bar{\alpha}_{1,2}$.

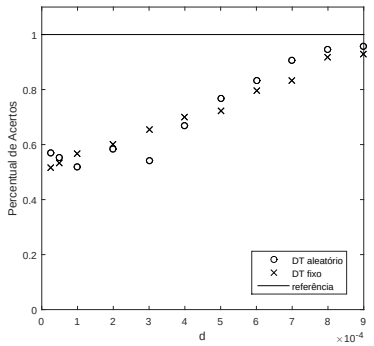

Figura 18: $\alpha = 10^{-5}$. Acertos da clusterização para centro aleatório e fixo.

3.3. Análise de resultados primeiro e segundo experimentos

No Primeiro Experimento, a escolha do domínio espacial na obtenção da solução numérica não gerou dados que representem bem o caso do coeficiente de difusão $\alpha = 1$. Já no Segundo Experimento, o domínio espacial utilizado para a geração dos dados melhorou os resultados para coeficientes $\alpha = 1$ a $\alpha = 10^{-2}$.

Em ambos os experimentos, coeficientes da ordem de 10^{-3} a 10^{-5} foram subestimados. É relevante ressaltar que apesar dos coeficientes recuperados para as ordens 10^{-3} a 10^{-5} serem subestimados após a separação, eles já haviam sido subestimados do ponto de vista de um foco.

Desta forma, observa-se que mesmo nesses casos o tratamento de separação de dados foi relevante, ou seja, os resultados foram condizentes com o esperado pelo método. Também, para esses valores de α, o tratamento de separação de dados não foi relevante.

Após a finalização deste trabalho, constatou-se que tal efeito pode ser corrigido aumentando-se a discretização da malha temporal e, também, aumentando-se a precisão computacional utilizada.

No Primeiro Experimento, a inicialização aleatória ou fixa do centro do cluster não teve influência significativa nos resultados. Também observou-se, em ambos os experimentos, que o tratamento de separação de dados foi relevante para os coeficientes de $\alpha = 1$ a $\alpha = 10^{-2}$.

Para o caso de menores ordens de grandeza, pode-se aumentar a precisão computacional e verificar se há uma melhora dos resultados. Outra possibilidade é realizar uma mudança de escala nos dados prévia à aplicação do algoritmo de clusterização.

No Segundo Experimento, o percentual de acertos do algoritmo de separação de dados foi maior com relação a uma mesma ordem de grandeza do coeficiente no Primeiro Experimento. Também observa-se que a escolha aleatória ou fixa do parâmetro DT não influenciou muito no percentual de acertos.

4. Aplicação prática: a gripe aviária na Nigéria

Em (Souza, 2010), estudou-se a dispersão do risco de contágio de pelo vírus H5N1, epidemia conhecida como gripe aviária. No referido trabalho, os

dados foram retirados de *World Organisation for Animal Health* (OIE, 2009) e foram feitas estimativas para o coeficiente de difusão de oito países.

A fim de analisar a aplicabilidade para problemas reais do método proposto por Souza (Souza, 2014) e, também, a relevância do tratamento de separação de dados, foi realizado experimento aqui apresentado.

No presente trabalho foi estudado apenas o resgistro de eventos para essa epidemia na Nigéria. Os dados de entrada são de latitude, longitude e dia de ocorrência. As autoridades de saúde, na época, identificaram a ocorrência de dois focos, motivando a escolha desse país para essa abordagem.

Observou-se dois eventos iniciais no dia 1 em locais distintos, um evento e_1 ocorrido nas coordenadas geográficas $(10,78\,;\,7,76)$ e outro evento e_2 ocorrido nas coordenadas geográficas $(11,66\,;\,8,53)$, o que poderia ser um indício de ocorrência de dois focos.

No entanto, tais eventos inicias estão muito próximos entre si, então é pouco provável que tais eventos sejam focos de surtos distintos. Caso os dois eventos iniciais e_1 e e_2 sejam classificados no mesmo *cluster*, é interessante calcular o coeficiente de difusão tomando como foco um e depois o outro, e comparar o resultado obtido.

A unidade de medida de tempo é em dias e as coordenadas geográficas foram transformadas em radianos.

Para o tratamento de separação de dados foram utilizados dois algoritmos de clusterização distintos: um K-means modificado e o algoritmo de Souza.

Para os dados originais, o coeficiente de difusão α obtido com relação ao foco e_1 foi $\alpha = 3,94\text{km}^2/\text{dia}$ e, tendo e_2 como foco, $\alpha = 4,49\text{km}^2/\text{dia}$.

4.1. Algoritmo de *Cluster* utilizado por Souza

Através do algoritmo de *Cluster* utilizado em (Souza, 2010), obtém-se a seguinte separação espacial para os dados.

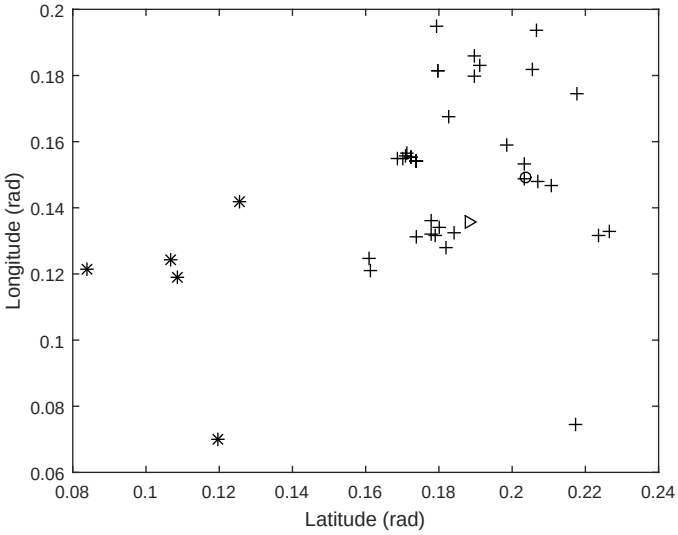

Figura 19: Gráfico latitude versus longitude em radianos para os *clusters* obtidos. O ponto referencial para o cálculo das distâncias é o círculo. Nesse caso, obteve-se $\alpha_1 = 1,54$ em $+$, considerando e_1 como foco e $\alpha_1 = 2,97$ com e_2. Para o outro *cluster* em $*$, $\alpha_2 = 0,66$.

No gráfico (19), para os dados separados, $\alpha_1 = 1,54$ considerando e_1 como foco e $\alpha_1 = 2,97$ considerando e_2. Para o outro *cluster* obtido (azul), $\alpha_2 = 0,66$. Percebe-se uma nítida separação espacial dos pontos, o que indica uma boa clusterização através desse algoritmo.

4.2. Algoritmo K-means adaptado para latitude e longitude

Através do algoritmo de *Cluster* K-means para latitude e longitude, obtém-se a seguinte separação espacial para os dados.

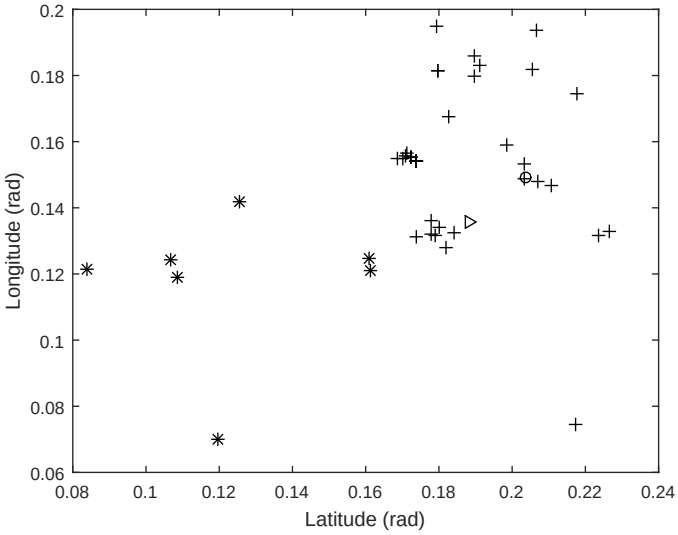

Figura 20: Gráfico latitude versus longitude em radianos para os *clusters* obtidos. Para o *cluster* em $+$, $\alpha_1 = 3,08$ considerando como foco e_1 e $\alpha_1 = 2,91$ considerando e_2. Para o *cluster* em $*$, $\alpha_2 = 0,97$.

No gráfico (20), os coeficientes de difusão obtidos após a separação foram $\alpha_1 = 3,08$ considerando como foco e_1 e $\alpha_1 = 2,91$ considerando e_2, em $+$. Para o outro *cluster*, em $*$, o resultado é $\alpha_2 = 0,97$. O coeficiente de difusão $\alpha_1 = 3,08$ é próximo do obtido para os dados originais, $\alpha = 3,94$.

4.3. Análise de resultados

A tabela (1) expõe os resultados para o cálculo dos coeficientes de difusão para os dados originais, em cada caso considerando um evento inicial como foco. Já a tabela (2), expõe o resultado para os dois algoritmos de clusterização considerados, o desenvolvido por Souza e o K-means modificado.

Tabela 1: Coeficientes de difusão α em km^2/dia para os dados originais considerando como foco os eventos iniciais e_1 e e_2.

Foco	α
e_1	$3,94$
e_2	$4,49$

Tabela 2: Coeficiente de difusão α_1 e α_2 em km^2/dia obtidos através de diferentes algoritmos de separação. Para α_1, foram considerados focos os eventos iniciais e_1 e e_2.

	Souza	**K-means**
α_1 (e_1)	$1,54$	$3,08$
α_1 (e_2)	$2,97$	$2,91$
α_2	$0,66$	$0,97$

A partir da tabela (1) observa-se que os coeficientes considerando diferentes eventos iniciais como foco são próximos, isto é, $\alpha = 3,94$ km^2/dia e $\alpha = 4,49$ km^2/dia.

Na tabela (2), para o algoritmo de Souza, para o cálculo de α_1, houve impacto da escolha dos eventos iniciais tomados como foco. Considerando e_1, $\alpha = 1,54$ km^2/dia e, para e_2, $\alpha = 2,97$ km^2/dia.

Agora, tendo em vista que a estimativa de tais coeficientes representam o espalhamento da mesma epidemia, num mesmo país, com condições geográficas parecidas, espera-se que os coeficientes para *clusters* distintos seja próximo.

Com base nessa hipótese, os coeficientes de *clusters* distintos que mais se aproximam são $\alpha_1 = 1,54$ km^2/dia e $\alpha_2 = 0,66$ km^2/dia. Desta forma, acredita-se que tais valores estejam mais próximos do espalhamento ocorrido no processo difusivo real.

Ainda com relação à tabela (2), para o algoritmo K-means modificado, houve menos impacto no cálculo de α_1 ao considerar os diferentes eventos iniciais como foco, $\alpha_1 = 3,08$ km^2/dia para e_1 e $\alpha_1 = 2,97$ km^2/dia para e_2.

Também considerando a hipótese de que coeficientes de diferentes *clusters* sejam próximos um do outro, os valores que possivelmente representem

melhor o espalhamento são $\alpha_1 = 2,91$ km^2/dia e $\alpha_2 = 0,97$ km^2/dia. Outra observação relevante é que o coeficiente $\alpha_1 = 3,08$ km^2/dia se aproxima do valor obtido para os dados originais, $\alpha = 3,94$ km^2/dia.

5. Conclusões

No presente trabalho, percebe-se que o tratamento de separação de dados pode ser imprescindível caso deseje-se obter resultados fidedignos para a estimativa do coeficiente de difusão. A verificação da relevância de tal tratamento foi feita através de dois experimentos distintos com relação à geração de dados.

A diferença entre os experimentos consistiu na escolha de modelagem do domínio espacial para a obtenção da solução numérica da concentração e posterior geração de eventos. Em ambos experimentos, a técnica empregada foi a geração de dois conjuntos de dados inicialmente disjuntos; em seguida, foram sobrepostos no mesmo domínio e, então, aplicado o tratamento de separação com relação ao atributo espacial.

Outra característica considerada foi a ordem de grandeza do coeficiente de difusão. Percebe-se que a separação de dados surtiu efeitos positivos para $1 \times 10^{-3} \leq \alpha \leq 1 \times 10^{0}$ no domínio espacial considerado. Em contrapartida, para coeficientes menores, houve uma subestimativa de tal parâmetro.

Tanto no Primeiro Experimento quanto no Segundo Experimento, o coeficiente de difusão já havia sido subestimado quando os dados foram vistos do ponto de vista de apenas um foco, em (Souza, 2014). Ao se aplicar a clusterização, tal característica foi ainda mais evidenciada, ou seja, obteve-se coeficientes de difusão ainda menores. Tal efeito pode ser observado devido à dificuldades na geração dos dados que representem bem tal ordem de grandeza do coeficiente de difusão.

Assim, nesse caso, mesmo que o método estime corretamente o coeficiente de difusão, os dados simulados talvez já não representem bem o processo difusivo com o coeficiente suposto. Com relação à clusterização, ainda com relação aos coeficientes de ordem menor do que 1×10^{-3}, a separação de dados pode não ser relevante.

Em muitos experimentos para coeficientes de tal ordem, a separação dos dados gerou *clusters* vazios à medida que a distância entre os focos diminuía. Tal resultado pode ser explicado por haver diferentes ordens de grandeza entre

o processo difusivo e o observador. No caso, o observador do processo é a máquina e quanto maior a precisão computacional, mais fidedígnos serão os resultados.

Desta forma, uma alternativa para se obter melhores resultados para coeficientes iguais ou menores do que 1×10^{-3} seria, em trabalhos futuros, impor maior precisão computacional. Outra possiblidade, para os casos em que a precisão da máquina não é suficiente para representar a ordem de grandeza trabalhada, seria realizar uma mudança de escala previamente à aplicação do algoritmo de clusterização.

Outro aspecto desse trabalho, é a aplicação da técnica de separação dos dados para a recuperação do coeficiente de difusão para o espalhamento do risco de contágio do vírus H5N1 na Nigéria. De acordo com as autoridades, haviam sido detectados, na época, dois focos para esse país e, por isso, escolheu-se separar os dados em duas partições.

Apesar de terem ocorridos dois eventos iniciais no dia 1 em locais distintos, a proximidade entre eles fez com que se descartasse a possibilidade de ambos serem focos de surtos distintos. Foram feitas estimativas para o coeficiente de difusão considerando cada um dos eventos iniciais. Para os dados originais, para o foco e_1, o coeficiente obtido foi $3,94$ km^2/dia, enquanto para a outro, $4,49$ km^2/dia.

Como o registro de eventos eram do tipo dia, latitude e longitude, foram utilizados dois algoritmos de clusterização que lidassem com esse tipo de entrada. O primeiro algoritmo, desenvolvido por Souza, resultou para um *cluster* valores $1,54$ km^2/dia e $2,97$ km^2/dia para focos distintos e, para o outro *cluster*, $0,66$ km^2/dia.

Já para o segundo algoritmo, o K-means modificado, para um *cluster* obtiveram-se $3,08$ km^2/dia e $2,91$ km^2/dia considerando cada um dos focos e, para o outro *cluster*, $0,97$ km^2/dia. Tendo em vista que a estimativa do coeficiente de difusão é referente ao espalhamento do mesmo vírus, situado num mesmo país, com condições geográficas semelhantes, espera-se que para cepas diferentes os parâmetros sejam próximos.

Considerando tal hipótese, observou-se que os resultados obtidos são coerentes: utilizando o algoritmo de Souza, $\alpha_1 = 1,54$ km^2/dia e $\alpha_2 = 0,66$ km^2/dia e, utilizando o K-menas modificado, $\alpha_1 = 2,91$ km^2/dia e $\alpha_2 = 0,97$ km^2/dia.

Com relação aos algoritmos de clusterização, observa-se que ambos uti-

lizam como atributo apenas a localização espacial dos eventos. Seria relevante, em trabalhos futuros, considerar também o impacto de se separar os eventos com relação à coordenada temporal para a estimativa do coeficiente de difusão.

Feitas tais considerações, conclui-se que os resultados desse trabalho corroboram com a relevância do tratamento de separação de dados para a estimativa do coeficiente de difusão, caso almeje-se atingir resultados fidedígnos.

Referências

Cao, Y. (2008). Efficient K-Means clustering using JIT. URL: http://www.mathworks.com.

CodeCodex (2015). Calculate distance between two points on a globe. URL: http://www.codecodex.com.

Gan, G., Ma, C., e Wu, J. (1979). *Data Clustering: Theory, Algorithms and Applications*. SIAM, Philadelphia.

Jain, A. K. . (2010). Data clustering: 50 years beyond k-meas. *Pattern Recognition Letters*, 31:651–666.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, páginas 281–297, Berkeley/CA, USA. University of California Press.

OIE (2009). The world organisation for animal health (oie). URL: http://www.oie.int.

Ross, S. M. (2006). *Simulation*. Elsevier, California.

Souza, J. M. R. (2010). Estudo da dispersão de risco de epizootias em animais: o caso da influenza aviária. Dissertação de Mestrado, IMECC–UNICAMP, Campinas/SP.

Souza, J. M. R. (2014). *Estimativa do Coeficiente de Difusão para Problemas (prioritariamente) Ecológicos*. Tese de Doutorado, IMECC–Unicamp, Campinas/SP.