Teoria da Informação Aplicada à Caracterização de Velocidades em Redes Veículares

Tamer S. G. Cavalcante[1], Andre L. L. Aquino[1], Osvaldo A. Rosso[1,2], Evellyn S. Cavalcante[3], and Eliana S. Almeida[1]

[1] LaCCAN/CPMAT – Instituto de Computação,
Universidade Federal de Alagoas (UFAL) – Maceió, AL – Brasil
[2] Laboratorio de Sistemas Complejos – Facultad de Ingeniería
Universidad de Buenos Aires (UBA) – Buenos Aires – Argentina
[3] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil
{tamersgc,alla.lins,oarosso,
evellyn.cavalcante,eliana.almeida}@gmail.com

Resumo This work presents a characterization of VANETs behaviour based on information theory. Among the vehicle information (displacement, time of trip, velocity etc), we concentrated in velocity one. To perform our study, the vehicle information was extracted from two datasets: Borlänge GPS and Mobile Century. The individual velocity of each vehicle was characterized as a time series; and the global velocities behaviour was characterized with the Bandt and Pompe methodology combined with causality Complexity-Entropy plane. Based on the corresponding localization, in the Complexity-Entropy plane, we could observe that the global velocity behaviour is compatible with a correlated noise f^{-k}.

Keywords: Vehicular networks. Information theory. Complexity-Entropy plane.

1 Introdução

Redes Veiculares Ad-hoc, (*Vehicular Ad-hoc Networks* – VANETs) [1] são redes móveis cujos nós são veículos equipados com tecnologia de comunicação sem fio. Esses veículos, além de interagirem entre si, podem se comunicar com a infraestrutura montada na região para transmitir e receber dados [2]. Os dados provenientes das VANETs podem fornecer informações sobre qualidade da estrada, congestionamentos e acidentes. Estas informações são manipuladas e utilizadas por diversas aplicações, como por exemplo, sistema de alerta de segurança, assistência ao motorista e roteamento de tráfego [3].

Devido às características estocásticas do tráfego de veículos, o entendimento do comportamento de tais características é uma tarefa desafiadora. O entendimento dessas características pode auxiliar na proposição de algoritmos que

atendam de forma mais assertiva as necessidades das aplicações que fazem uso dessas redes veiculares. Algumas características estão diretamente relacionadas ao comportamento do tráfego de veículos, por exemplo, a variabilidade da velocidade que, por sua vez, afeta a topologia e a conectividade da rede. A topologia numa VANET apresenta um comportamento dinâmico ocasionando desconexões e dificultando a transmissão de dados [1].

Uma ferramenta bastante utilizada para a modelagem de sistemas dinâmicos e que se ajusta ao cenário de redes veiculares são as séries temporais. As séries temporais são importantes, pois armazenam informações de eventos passados que, quando analisadas, podem ser úteis na realização de previsões. No nosso caso o sistema dinâmico considerado é a velocidade dos carros.

Estamos interessados em caracterizar o comportamento global das velocidades, ou seja, analisar conjuntamente as velocidades (séries temporais) de um conjunto de carros que trafegam em uma mesma região. Dessa forma, para medir a quantidade de informação contida nas velocidades, optamos por utilizar conceitos da Teoria da Informação, no nosso caso a entropia e complexidade estatística. De forma geral, a entropia pode ser definida como sendo o grau de "desordem" do sistema. Já a complexidade estatística é uma medida utilizada para avaliar o grau das estruturas ou os padrões de um processo. Com o objetivo de distinguir o determinismo e a aleatoriedade das velocidades, consideramos o plano Complexidade-Entropia, que mostra o grau de desordem do sistema com base no graus das estruturas.

Para os testes realizados extraímos as informação das velocidades dos veículos de duas bases de dados: Borlänge GPS [4] [5] e *Mobile Century* [6]. Observando nos resultados a localização das velocidades no plano, e verificamos que o comportamento global das velocidades é compatível com o ruído correlacionado f^{-k} [7]. A identificação desse comportamento nos ajuda, por exemplo, na proposição de algoritmos de disseminação de dados, que, nesse caso, devem favorecer a propagação dos dados entre os veículos considerando um comportamento global próximo a um ruído f^{-k}.

O trabalho está organizado como segue: A Seção 2 apresenta alguns trabalhos relacionados ao nosso estudo; A Seção 3 mostra a metodologia para a caracterização das velocidade; A Seção 4 descreve os principais resultados. A Seção 5 mostra as considerações finais e as propostas para trabalhos futuros.

2 Trabalhos relacionados

De forma geral, muitos trabalhos utilizam a metodologia de caracterização de séries temporais considerando o plano Complexidade-Entropia. A título de exemplo, Zunino et al. [8] utilizam o plano para avaliação do mercado de ações, nesse caso a variação das ações são modeladas como séries temporais. Eles apresentam uma aplicabilidade satisfatória para os mecanismos acima apresentados, pois conseguem distinguir os estágios de desenvolvimento do mercado (emergentes, desenvolvidos e um grupo intermediário). As diferenças entre os estágios são facilmente visualizadas quando as informações são inseridas no plano.

Buscando trabalhos que consideram o estudo de velocidades de redes veiculares que utilizam o plano Complexidade-Entropia encontramos os trabalhos de Shang et al. [9] e Liao et al. [10]. Shang et al. [9] que analisa a velocidade dos veículos utilizando a base de dados *Beijing Yuquanying*. As informações de velocidade são adquiridas por intermédio de uma infraestrutura montada nas rodovias, que coletam os dados a cada 20 segundos. No entanto, nosso caso, consideramos uma base de dados onde os próprios carros são responsáveis por enviar suas informações (GPS). Dados coletados por GPS são mais susceptíveis a erros, logo exige um tratamento mais acurado dos dados e, consequentemente, métodos mais robustos para sua análise.

De forma complementar, porém com outro objetivo, Liao et al. [10] analisaram o índice de congestionamento de tráfego da cidade de Pequim através do plano Complexidade-Entropia. Tal índice é calculado com base nas velocidades dos carros e representa o status do tráfego, ou seja, se está congestionado ou fluindo. Os dados foram coletados através de informações locais pelas as ruas da cidade. Os resultados mostraram que o plano é o melhor modo de classificar os níveis de congestionamento em relação a outros métodos apresentados no trabalho.

Como a utilização do plano Complexidade-Entropia não é a única forma de caracterizar as velocidades em uma VANET, Shang et al. [9], além do plano Complexidade-Entropia, também utilizaram a dimensão fractal aplicada por intermédio da modelagem multifractal. Os resultados mostraram que o trânsito possui características multifractais e que o grau de fractalidade do tráfego tende a aumentar à medida que o congestionamento aumenta. Além disso, o expoente Hölder [11] é proposto como indicador para prever a presença de congestionamentos, pois ele mede a taxa local de fractalidade. No nosso trabalho, por outro lado, utilizamos a entropia e outros conceitos da teoria da informação para analisar os dados nos permitindo assim uma análise mais detalhada do comportamento global dos veículos.

3 Identificação do comportamento das VANETs

A principal motivação para a proposição do estudo sobre o comportamento das VANETs é derivada do seguinte problema:

> *Como caracterizar o tráfego de veículos considerando apenas suas velocidades?*

Para apresentar uma resposta factível para essa pergunta modelamos a velocidade de cada veículo como uma série temporal e aplicamos técnicas de Teoria da Informação. Com isso, para medir a quantidade de informação agregada de cada veículo em relação ao comportamento global do tráfego utilizamos o plano Complexidade-Entropia [7].

Para determinarmos o plano Complexidade-Entropia são necessários os seguintes passos: (i) determinar a *Função Distribuição de Probabilidade*; (ii) calcular a *Entropia de Shannon Normalizada*; e (iii) calcular a *Complexidade Estatística*, que são discutidos e detalhados nas próximas seções.

3.1 Função de distribuição de probabilidade

De forma geral, o *Método Bandt e Pompe* é utilizado para estimar a *Função de Distribuição de Probabilidade* P de uma série temporal que no nosso caso são as velocidades dos veículos. Para essa estimação, o método, leva em conta a causalidade, comparando os valores correntes com os dos vizinhos na série temporal. Sendo um pouco mais formal, para cada tempo s da série temporal $\mathcal{X} = \{x_t : t = 1, 2, \ldots, M\}$, um vetor de tamanho D é construído da seguinte forma:

$$(s) \mapsto (x_{s-(D-1)}, x_{s-(D-2)}, \ldots, x_{s-1}, x_s).$$

O valor D vem de dimensões embutidas, que determina a quantidade de informação contida em cada vetor a fim de revelar detalhes importantes da estrutura.

Em seguida, para cada vetor da série temporal, um padrão ordinal é associado. O padrão ordinal é definido como a permutação $\pi = (r_0 r_1 \ldots r_{D-1})$ de $(0, 1, \ldots, D-1)$ definido como:

$$x_{s-r_{D-1}} \leq x_{s-r_{D-2}} \leq \cdots \leq x_{s-r_1} \leq x_{s-r_0}.$$

Ou seja, o vetor será ordenado de forma crescente e a permutação π será criada de acordo com a mudança dos valores permutados.

Então, para todas as permutações de π, a função de distribuição de probabilidade $P = \{p(\pi)\}$ é definida por

$$p(\pi) = \frac{\#\{s | s \leq M - D + 1; (s), \text{tem o padrão } \pi\}}{M - D + 1}. \tag{1}$$

Como ilustração da aplicação do método, seja $\mathcal{X} = (50, 20, 100, 30, 60, 113, 114)$ a série temporal que representa um conjunto de velocidades medidas. Para $D = 3$, o vetor de valores correspondente a $s = 1$ será $(50, 20, 100)$. Ao ordená-lo o vetor ficará desta forma: $(20, 50, 100)$ que corresponde à permutação $\pi = (1, 0, 2)$. Para $s = 2$, tem-se o vetor $(20, 100, 30)$. Daí, ao ordená-lo ficará $(20, 30, 100)$ que refere-se a permutação $\pi = (0, 2, 1)$ e assim por diante. Ao fim, temos a sequência de ocorrência de todos os padrões. Tal sequência dividida pela quantidade total, conforme descrito na Equação 1, corresponderá à distribuição de probabilidades da série $\mathcal{X}$.

3.2 Entropia de Shannon Normalizada

Para Shannon [12] a entropia é uma medida de incerteza de um processo físico descrito por uma *Função de Distribuição de Probabilidade* $P = \{p_i : 1, 2, \ldots, N\}$. A *Entropia de Shannon* é definida por

$$S[P] = -\sum_{i=1}^{N} p_i \ln p_i. \tag{2}$$

Se $S[P] = S_{min} = 0$, então é possível prever que a saída i, cuja probabilidade é p_i, realmente vai ocorrer. Ou seja, possuí-se um conhecimento máximo sobre o sistema. Por outro lado, o conhecimento é mínimo quando se trata de sistemas cujo comportamento é descrito pela distribuição uniforme, cuja probabilidade é dada por $P_e = \{1/N : i = 1, 2, \ldots, N\}$. Logo, $S_{max} = S[P_e] = \ln N$.

No entanto, utilizamos a entropia normalizada conforme definida Martin [13]. Assim, seja P uma *Função de Distribuição de Probabilidade* e $S[P]$ a *Entropia de Shannon* definida na equação 2, temos que a *Entropia de Shannon Normalizada* $\mathcal{H}$ é dada por

$$\mathcal{H}[P] = \frac{S[P]}{S_{max}}.$$

A utilização da *Entropia de Shannon Normalizada* para estimar a distribuição de probabilidade oriunda do *Método Bandt e Pompe*, como fazemos nesse trabalho, é chamada de *Entropia de Permutação Normalizada* [14] que é definida como

$$\mathcal{H}[P] = -\frac{1}{\ln D!} \sum_{i=1}^{D!} p(\pi_i) \ln p_i(\pi). \tag{3}$$

Lembrando que devemos ter sempre $M >> D!$ e $D! = N$.

3.3 Complexidade Estatística

Complexidade Estatística é uma medida utilizada para avaliar o grau das estruturas ou os padrões de um processo. Ela é dada a partir da escolha de uma distância entre uma *Função de Distribuição de Probabilidade* P e uma distribuição de referência P_e, que também pode se chamar de desequilíbrio. Podemos definir *Complexidade Estatística* [15] [16] como

$$\mathcal{C}[P] = \mathcal{Q}[P, P_e] \cdot \mathcal{H}[P],$$

onde $\mathcal{Q}$ é o desequilíbrio e $\mathcal{H}$ é a *Entropia de Shannon Normalizada*, que no nosso trabalho é a *Entropia de Permutação Normalizada*.

O desequilíbrio $\mathcal{Q}$ deve refletir a "arquitetura" de um sistema, sendo diferente de zero se houver alguma estrutura que seja "privilegiada", ou estados que sejam mais prováveis entre os acessíveis. Ele é definido em termos da *Divergência de Jensen-Shannon* $\mathcal{J}[P, P_e]$, cuja escolha é discutida em [17]. Então,

$$\mathcal{Q}[P, P_e] = \mathcal{Q}_0 \cdot \mathcal{J}[P, P_e],$$

onde,

$$\mathcal{J}[P, P_e] = S\left[\frac{(P + P_e)}{2}\right] - \frac{S[P]}{2} - \frac{S[P_e]}{2},$$

e $\mathcal{Q}_0$ é uma constante de normalização, logo $0 \leq \mathcal{Q} \leq 1$. O valor de $\mathcal{Q}_0$ é igual ao inverso do valor máximo possível $\mathcal{J}[P, P_e]$, que é dado por

$$\mathcal{Q}_0 = -2\left\{\left(\frac{N+1}{N}\right)\ln(N+1) - 2\ln 2N + \ln N\right\}^{-1}.$$

3.4 Plano Complexidade-Entropia

Os quantificadores, baseado em Teoria da Informação, definidos anteriormente $\mathcal{H}$ e $\mathcal{C}$ avaliados com a *Função de Distribuição de Probabilidade* oriunda do *Método Bandt e Pompe*, nos permite definir o plano $\mathcal{H} \times \mathcal{C}$. O *Plano Complexidade-Entropia*, $\mathcal{H} \times \mathcal{C}$, é baseado apenas nas características globais da *Função de Distribuição de Probabilidade* associada e ambos os quantificadores são definidos em termos da *Entropia de Shannon*. O intervalo de variação é $[0, 1] \times [C_{min}, C_{max}]$, onde C_{min} e C_{max} são os valores da *Complexidade Estatística* mínima e máxima, respectivamente, para um dado valor de $\mathcal{H}$.

O *Plano Complexidade-Entropia* é uma ferramenta que tem sido utilizada para tratar o determinismo e a aleatoriedade em séries temporais em geral. Rosso et al. [7] utilizou o plano para distinguir os sistemas caóticos dos processos estocásticos, ou seja, separar o que é caos do que é ruído. O uso do plano para a caracterização das velocidades possibilita a compreensão das caraterísticas dinâmicas e melhora o entendimento sobre o conjunto de dados estudado.

4 Resultados e Discussões

Para podermos identificar o comportamento das VANETs baseado em suas velocidades utilizamos duas bases de dados, que são descritas na Seção 4.1. Em seguida efetuamos os devidos tratamentos para as velocidades de ambas as base (Seção 4.2). Por fim, apresentamos a representação do plano (Seção 4.3).

4.1 Descrição das Bases de Dados

Borlänge GPS: A base de dados representa a cidade de Borlänge que possui 3.077 cruzamentos interconectados por 7.459 estradas. A coleta dos dados ocorreu durante dois anos, para um estudo de segurança no trânsito, e envolveu quase 200 veículos. Os veículos foram equipados por um GPS e monitorados num raio de 25 km ao redor do centro da cidade e sua posição gravada. A parcela da base disponível para utilização, no entanto, possui dados de apenas 24 veículos, correspondendo a 420.814 observações. Para maiores informações, ver [4] [5].

Na base de dados as informações de tempo e deslocamento não estão explícitas, mas é possível obtê-las através da manipulação dos três arquivos que compõe a base: `mobility`, `nodepos` e `nodes`. O arquivo `mobility` (arquivo 1.1) possui as informações do identificador do veículo, dia da viagem, número da viagem, hora inicial e final do percurso, dentre outras.

Arquivo 1.1. mobility

```
1 4,1,2,2.42,1,1,0.016000,0.002000,2000-11-10 14:24:11,2000-11-10 14:24:19,
2 4,1,2,2.42,2,1,0.002000,0.153821,2000-11-10 14:24:19,2000-11-10 14:24:33,
3 4,1,2,2.42,3,1,0.000000,0.074735,2000-11-10 14:24:33,2000-11-10 14:24:59,
4 4,1,2,2.42,4,3,0.219877,0.000000,2000-11-10 14:24:59,2000-11-10 14:25:18,
5 4,1,2,2.42,5,1,0.006453,0.000000,2000-11-10 14:25:18,2000-11-10 14:25:23,
6 4,1,2,2.42,6,1,0.174354,0.000000,2000-11-10 14:25:23,2000-11-10 14:26:17,
7 4,1,2,2.42,7,1,0.000000,0.063000,2000-11-10 14:26:17,2000-11-10 14:26:32,
8 4,1,2,2.42,8,3,0.063000,0.000000,2000-11-10 14:26:32,2000-11-10 14:26:32,
9 4,1,2,2.42,9,1,0.000000,0.039000,2000-11-10 14:26:32,2000-11-10 14:26:36,
```

No exemplo acima (arquivo 1.1), a primeira coluna se refere à identificação do veículo 4. A base disponibiliza informações de mais 23 veículos. A segunda coluna está relacionada as informações do dia 1 do veículo 4. Esse veículo ainda possui registro de mais 150 dias. A coluna seguinte, apresenta as viagens dos veículos, no exemplo, a viagem 2 do veículo 4 no dia 1. Por fim, a última e penúltima coluna são, respectivamente, a hora inicial e final do percurso. As outras colunas não são utilizadas no trabalho e foram ignoradas.

Arquivo 1.2. nodes

```
 316 1076
1076 316
 316 792
 792 2611
2611 321
 321 1823
1823 318
 318 1823
1823 321
```

Arquivo 1.3. nodepos

```
 316 , 15.443687 , 60.476045
1076 , 15.445492 , 60.474991
 792 , 15.44258 , 60.475656
2611 , 15.440816 , 60.477419
 321 , 15.440701 , 60.47741
1823 , 15.438019 , 60.476698
 318 , 15.44126 , 60.475159
 ...
 ...
```

O arquivo `nodes` (arquivo 1.2) possui o mesmo número de linhas do arquivo `mobility` e em cada uma delas possui dois valores, que referem-se ao ponto inicial e final de um percurso. Esses pontos são duas coordenadas, latitude e longitude, que estão no arquivo `nodepos` (arquivo 1.3).

Combinando todos os arquivos (arquivos 1.1-1.3), o veículo 4, no dia 1 e na viagem 2, viajou do ponto 316 (latitude = 60.476045 e longitude = 15.443687) para o ponto 1076 (latitude = 60.474991 e longitude = 15.445492), saindo às 14:24:11 do dia 10/11/2000 e chegando às 14:24:19 do mesmo dia.

***Mobile Century*:** Os dados foram coletados durante o experimento *Mobile Century* em 08 de Fevereiro de 2008 entre às 10:00am e 18:00pm na rodovia interestadual 880 na Califórnia. A base de dados é composta de 77 logs de GPS individuais que foram extraídos de dispositivos móveis Nokia N95. As informações disponíveis para cada veículo são: tempo, as coordenadas de latitude e longitude e a velocidade em milhas por hora. Para maiores informações, ver Herrera et al.[6].

O arquivo 1.4 apresenta um trecho do log de um veículo da base. Na primeira coluna tem-se o *Unix time* em milissegundos. A segunda e a terceira coluna são as coordenadas, latitude e longitude, respectivamente. E por último, tem-se a velocidade do veículo em milhas por hora. Em média, os dados são coletados à cada 3s, podendo sofreram pequenas variações nesse tempo.

Arquivo 1.4. mobile

```
1202497202837,37.6004328519,-122.0637571325,0.009
1202497206836,37.6004329358,-122.0637568810,0.010
1202497209837,37.6004329358,-122.0637566295,0.013
1202497212837,37.6004329358,-122.0637563781,0.015
1202497216826,37.6004330196,-122.0637560428,0.016
1202497220826,37.6004331872,-122.0637556237,0.017
```

4.2 Tratamento dos Dados

Para utilizar a metodologia apresentada na Seção 3 foi necessário o tratamento da base *Börlange* e pequenas alterações na base *Mobile Century*. O objetivo desse tratamento é disponibilizar as informações das velocidades com o menor número de erros possível.

Base de Dados de Borlänge: Num primeiro momento, assumindo que a base de dados estão com todos os dados corretos, precisamos efetuar o calculo das velocidades para cada viagem de cada carro. Assim efetuamos o calculo da velocidade média, $v = \Delta_s/\Delta_t$, onde Δ_s é o deslocamento do veículo e Δ_t é a variação do tempo. No entanto, para obtermos o valor para o Δ_s em m, precisamos calcular a menor distância entre dois pontos (latitude, longitude). Para isso, utilizamos o *software* estatístico `R` [18], mais especificamento o pacote `geosphere` e a função `distMeeus()`, que calcula a menor distância entre dois pontos (latitude, longitude) de acordo com o método proposto por [19].

A Tabela 1 apresenta os valores de tempo, deslocamento e velocidade do exemplo da Seção anterior que combina todos os arquivos (arquivos 1.1-1.3) da base de dados.

Tabela 1. Valores de tempo, deslocamento e velocidade para um trecho da base de dados.

linha	arquivo	tempo (s)	deslocamento (m)	velocidade (m/s)
1	08	229.53656	28.692070
2	14	229.53656	16.395469
3	26	129.41146	04.977364
4	19	271.84871	14.307827
5	05	012.76163	02.552327
6	54	306.45007	05.675001
7	15	394.83424	26.322283
8	00	300.00000	NaN
9	04	306.45007	76.612518

Nesse exemplo, podemos identificar as principais inconsistências na base. Inicialmente, na linha 8 o tempo para percorrer 300 m é 0 s gerando uma divisão por zero. Em seguida, na linha 9 o veículo percorreu uma distância de aproximadamente 306 m em apenas 4 s, atingindo uma velocidade média de aproximadamente 76 ms (273 km/h).

Para o problema da divisão por zero, descartamos todos os valores que apresentaram esse problema. Já para os valores anormais de velocidades utilizamos um tratamento mais acurado. Inicialmente, usamos o *Boxplot* de todas as velocidades de todos os veículos para podermos identificar a presença de *outliers*, verificamos que existiam velocidades com valores de aproximadamente 6.000 m/s (21.600 km/h).

Para não fazermos o descarte arbitrário dos *outliers* efetuamos uma sequência de análises das velocidades para assim realizarmos o descarte. Inicialmente, calculamos a média das velocidades de cada viagem por veículo. Usamos o *Boxplot* dessas médias para identificar as viagens *outliers*, ou seja, as viagens que possuem, na sua maioria valores de velocidades discrepantes. Foram descartadas as viagens que estavam fora do intervalo interquartil. A Figura 1 ilustra o *Boxplot* com os *outliers* que foram descartados.

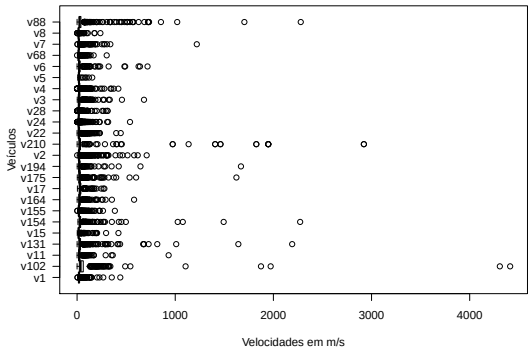


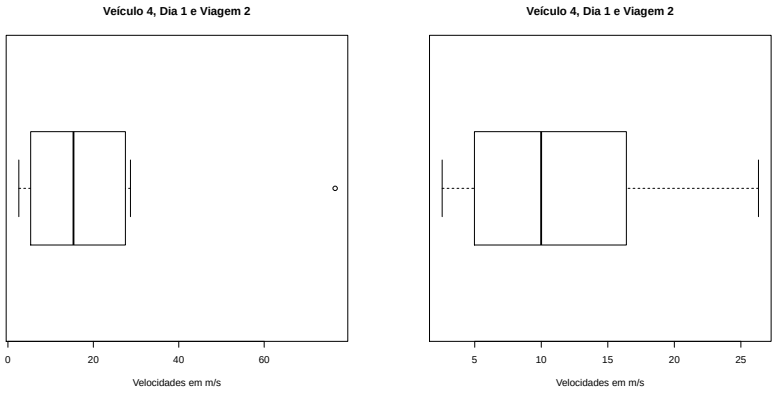
Figura 1. Boxplot das médias das viagens

Em seguida, para cada viagem não descartada, observou-se ainda existiam velocidades com comportamento anômalo. Com isso, para cada viagem usamos o *Boxplot* para descartar as velocidades *outliers* acima do quartil superior. Considerando a viagem do exemplo apresentado nos arquivos 1.1-1.3, as velocidades foram descartadas de acordo com o *Boxplot* apresentado na Figura 2.

Na Figura 2(a) observa-se a presença de um *outlier* e que foi descartado. Na Figura 2(b) apresenta o *Boxplot* da viagem sem a presença do *outlier*. Resgatando a Tabela 1, as velocidades essa viagem sofreram pequenas variações, porém na última linha tem um pico fora do comum, atingindo a velocidade médias de 76 m/s (273 km/h) num curto espaço de tempo.

Após isso, calculamos o *Boxplot* de todas as velocidades novamente para identificar se ainda existiam *outliers*. Identificamos que mesmo depois de todos os descartes realizados, ainda aparecem alguns *outliers*. Portanto, decidimos por descartar todas as velocidades que apresentavam outliers acima de 60 m/s (216 km/h). Esse descarte foi baseado na velocidade limite dos carros convencionais atuais, que atingem no máximo uma velocidade de aproximadamente 220 km/h.

Base de Dados *Mobile Century*: Diferentemente da base de dados de Borlänge, os dados do *Mobile Century* apresentam menos problemas em relação as veloci-



(a) Antes do descarte (b) Depois do descarte

Figura 2. Boxplot da viagem apresenta nos trechos do arquivo antes e depois do descarte.

dades. A Figura 3 mostra o Boxplot das velocidades dos veículos da base. Note que, não existem *outliers* e todas as velocidades são atingíveis por um carro convencional. Porém, foi preciso converter as velocidades dessa base para metros por segundo, pois os dados são em milhas por hora (mph). Então, para fazer essa conversão basta multiplicar o valor em mph por 0.44704.

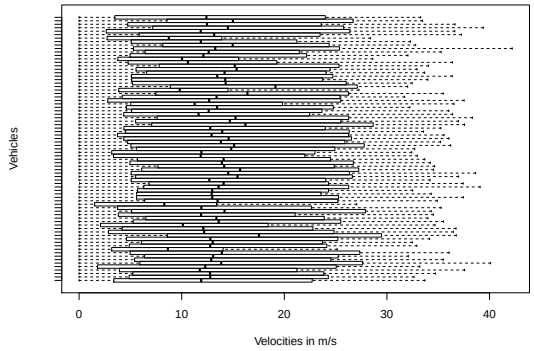
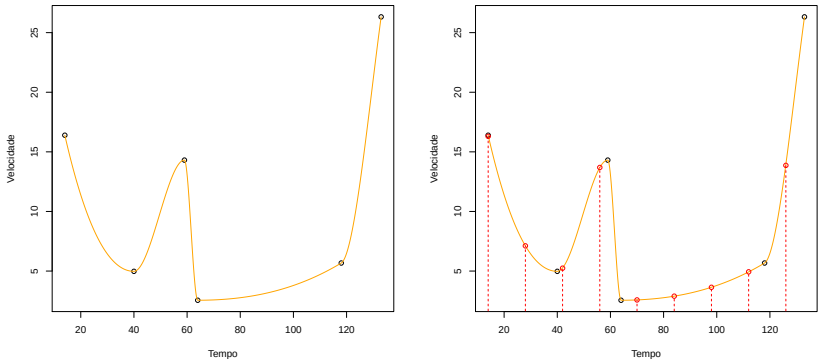


Figura 3. Boxplot das velocidades da base de dados do *Mobile Century*.

4.3 Caracterização das Velocidades

Para a caracterização das velocidades as mesmas precisam ser equidistantes, ou seja, o intervalo da leitura de todas as velocidade precisa ser o mesmo. Essa restrição é uma imposição do método utilizado [16]. Para isso, com as velocidades obtidas na Seção anterior, foi realizada uma interpolação, para que fosse possível obter valores realistas e equidistantes. A técnica de interpolação utilizada foi a *Piecewise Cubic Hermite Interpolating Polynomial* [20].

Considerando a viagem do exemplo da base de dados de Borlänge, apresentado nos arquivos 1.1-1.3, a Figura 4(a) apresenta o resultado da interpolação dos valores das velocidades da Tabela 1. As novas velocidades foram obtidas em um intervalo constante de 14 s. Esse tempo foi escolhido pois era o menor intervalo identificado dentre todas as viagens válidas da base de dados. A Figura 4(b) mostra como ficou a distribuição das velocidades do exemplo da Tabela 1 depois da interpolação dos dados. Os pontos em vermelho representam as velocidades obtidas.

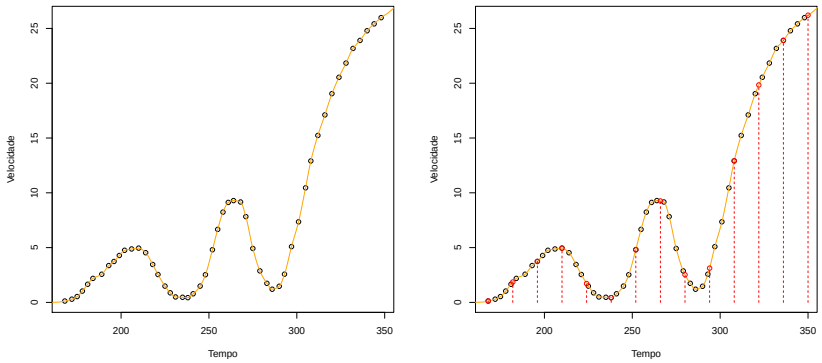


(a) Interpolação (b) Novas velocidades

Figura 4. Interpolação dos dados e as novas velocidades para a base de dados de Borlänge.

A composição da série temporal das velocidades de todas viagens interpoladas de todos os veículos foi realizada considerando o período em que cada uma delas ocorreram, por exemplo: dia 1, viagem 1; dia 1, viagem 2; dia 2, viagem 1; e assim por diante. O tempo do veículo parado entre cada viagem foi desconsiderado.

O mesmo foi feito para a base *Mobile Century*. Como os dados da base foram coletados num intervalo de 3s, então, da mesma como para Borlänge, nós interpolamos as velocidades e calculamos num intervalo de 14 segundos. A Figura 5 apresenta uma amostra das velocidades de um veículo da base. Observe que, muitas velocidades acabaram sendo descartadas por causa do pequeno intervalo que elas foram adquiridas.



(a) Interpolação (b) Novas velocidades

Figura 5. Interpolação dos dados e as novas velocidades para a base de dados *Mobile Century*.

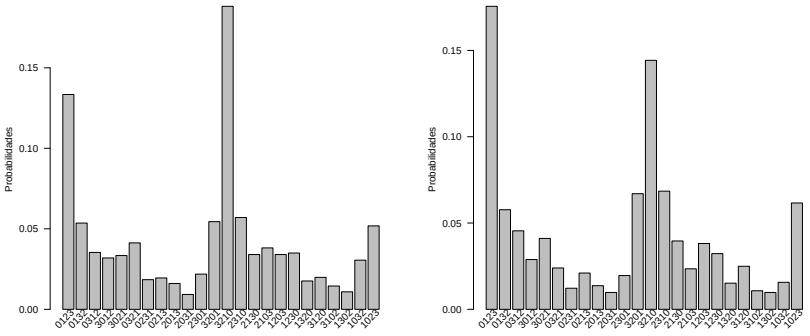
Utilizando a caracterização, descrita na Seção 3, para cada série temporal atribuímos uma função de distribuição de probabilidade através do *Método de Bandt e Pompe*, Seção 3.1. Esse método leva em consideração a causalidade, comparando os valores correntes com os vizinhos na série temporal de acordo com quantidade de dimensões embutidas. No nosso caso, utilizamos número de dimensões embutidas igual a 4. Na Figura 6 apresenta a função de distribuição de probabilidade para um veículo da base de Borlänge e outro da base *Mobile Century*.

Note que, na Figura 6 apesar dos veículos serem de bases totalmente diferentes, inclusive de outros país, ambos apresentam um comportamento semelhante. As probabilidades mais altas são do carro desacelerar e acelerar, ou seja, os padrões mais presentes são: 0123 e 3210.

Em seguida, calculamos: a *Entropia de Shannon Normalizada*, Seção 3.2, que define o grau de "desordem" dos dados; e a *Complexidade Estatística*, Seção 3.3, que avalia o grau das estruturas ou os padrões dos dados, para cada função de distribuição. De posse desses valores obtivemos o *Plano Complexidade-Entropia*, Seção 3.4, para assim realizamos a analise das séries temporais.

A Figura 7 apresenta o *Plano Complexidade-Entropia* com dimensões embutidas igual a 4 ($D = 4$). Plotamos no plano os dados extraídos, processados e interpolados da base de dados *Borlänge*, da base *Mobile Century* e a curva que representa o *K-noise* [21] [22]. Cada ponto no plano representa o comportamento das velocidades de um cada veículo de *Mobile Century*, cada cruz representa um veículo de Borlänge e cada triangulo um valor para o *K-noise*. A linha tracejada serve apenas para facilitar a visualização dos dados.

De acordo com a localização das velocidades dos veículos no plano, observamos que o comportamento das velocidades se assemelha ao do ruído. Para



(a) Borlänge (b) *Mobile Century*

Figura 6. Função de distribuição de probabilidade para um veículo da base de Borlänge e outro da base *Mobile Century*.

Borlänge o ruído varia entre $1.5 \leq K \leq 2.0$ e para *Mobile Century* a maioria dos valores variam entre $2.0 \leq K \leq 2.5$. Portanto, podemos observar que o comportamento global das velocidades para ambas bases de dados é compatível com o ruído correlacionado f^{-k}.

5 Considerações Finais

Este trabalho propôs um estudo sobre o comportamento da velocidade de veículos visando a melhoria no projeto das aplicação que executam nas VANETs. A caracterização foi realizada por intermédio de conceitos provenientes da Teoria da Informação: entropia e complexidade. O *Método de Bandt e Pompe* foi utilizado para atribuir uma *Função de Distribuição de Probabilidade* para às séries temporais que descrevem as velocidades dos veículos.

As informações dos veículos foram extraídas de duas bases de dados: *Borlänge GPS* e *Mobile Century*. Na base de dados *Borlänge*, ao realizar o cálculo da velocidade para cada carro, observou-se que alguns resultados não condiziam com a realidade, sendo necessário um pré-processamento. Já na *Mobile Century* só tivemos que fazer a conversão de milhas por horas para metros por segundo.

Plotamos a complexidade e a entropia das velocidades dos veículos, juntamente com os valores do ruído K. De acordo com a representação no plano, observamos que o comportamento das velocidades se assemelha ao do ruído colorido com K variando entre $1.5 \leq K \leq 2.5$. Portanto, podemos inferir que o comportamento global das velocidades é compatível com tal ruído.

Como trabalho futuro, aplicaremos a nossa caracterização em outras bases de dados para corroborar ainda mais nossos resultados. Além disso, fazer uma análise de como os ruídos coloridos refletem no comportamento do tráfego. Outro

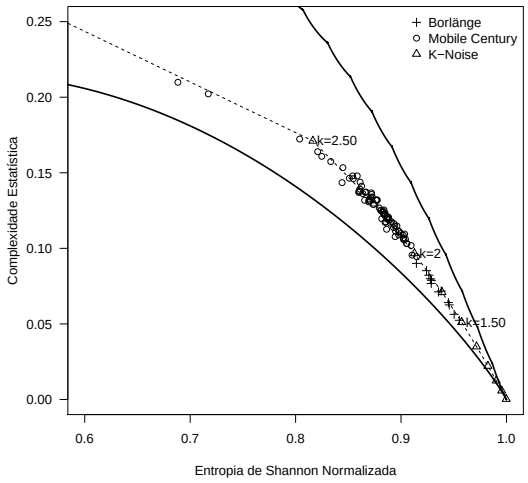


Figura 7. *Plano Complexidade-Entropia* com os valores da complexidade e entropia da base de dados *Borlänge* e do *K-noise*.

ponto importante é a utilização dessa informação, de que a velocidade é compatível com o ruído correlacionado, para auxiliar a proposição de novo algoritmos em redes veiculares, por exemplo, predição de trajetórias.

Referências

1. Yousefi, S., Mousavi, M., Fathy, M.: Vehicular ad hoc networks (VANETs): Challenges and perspectives. In: 6th International Conference on ITS Telecommunications Proceedings. (2006)
2. Hartenstein, H., Laberteaux, K.P.: A tutorial survey on vehicular ad hoc networks. Communications Magazine, IEEE **46**(6) (jun 2008) 164–171
3. Faezipour, M., Nourani, M., Saeed, A., Addepalli, S.: Progress and challenges in intelligent vehicle area networks. Communications of the ACM **55**(2) (February 2012) 90–100
4. Frejinger, E.: Route choice analysis. PhD thesis, SB, Lausanne (2008)
5. Freudiger, J., Shokri, R., Hubaux, J.: Evaluating the Privacy Risk of Location-Based Services. In: Financial Cryptography and Data Security. (2011)
6. Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q., Bayen, A.M.: Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transportation Research Part C: Emerging Technologies **18**(4) (2010) 568–583
7. Rosso, O.A., Larrondo, H.A., Martin, M.T., Plastino, A., Fuentes, M.A.: Distinguishing noise from chaos. Physical Review Letters **99**(15) (Oct 2007)
8. Zunino, L., Zanin, M., Tabak, B.M., Pérez, D.G., Rosso, O.A.: Complexity-entropy causality plane: A useful approach to quantify the stock market inefficiency. Physica A: Statistical Mechanics and its Applications **389**(9) (2010) 1891–1901

9. Shang, P., Lu, Y., Kama, S.: The application of Hölder exponent to traffic congestion warning. Physica A: Statistical Mechanics and its Applications **370**(2) (2006) 769–776
10. Liao, G., Shang, P.: Scaling and complexity-entropy analysis in discriminating traffic dynamics. Fractals **20**(03n04) (2012) 233–243
11. Daoudi, K., Vehel, J.L., Meyer, Y.: Construction of continuous functions with prescribed local regularity. Constructive Approximation **14**(3) (1998) 349–385
12. Shannon, C.E.: A Mathematical Theory of Communication. Bell system technical journal **27** (1948)
13. Martin, M., Plastino, A., Rosso, O.: Generalized statistical complexity measures: Geometrical and analytical properties. Physica A: Statistical Mechanics and its Applications **369**(2) (2006) 439 – 462
14. Bandt, C., Pompe, B.: Permutation Entropy: A Natural Complexity Measure for Time Series. Physical Review Letters **88**(17) (Apr 2002) 174102–174106
15. Ruiz, R.L., Mancini, H.L., Calbet, X.: A statistical measure of complexity. Physics Letters A **209**(5-6) (December 1995) 321–326
16. Rosso, O.A., De Micco, L., Larrondo, H.A., Martín, M.T., Plastino, A.: Generalized Statistical Complexity Measure. International Journal of Bifurcation and Chaos **20**(3) (2010)
17. Lamberti, P., Martin, M., Plastino, A., Rosso, O.: Intensive entropic non-triviality measure. Physica A: Statistical Mechanics and its Applications **334**(1) (2004) 119–131
18. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2012)
19. Meeus, J.: Astronomical Algorithms. Willman-Bell Inc. (1999)
20. Deboor, C.: A Practical Guide to Splines. Springer Verlag (1978)
21. Jakeman, E., Pusey, P.N.: A model for non-Rayleigh sea echo. IEEE Transactions on Antennas and Propagation **24**(6) (November 1976) 806–814
22. Jakeman, E.: On the statistics of K-distributed noise. J. Phys. A.: Math. Gen. **13**(1) (1980) 31–48