INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


# VANTAGENS E DESAFIOS DA UTILIZAÇÃO DE *WEB SCRAPING* NO INQUÉRITO AO TRANSPORTE RODOVIÁRIO DE MERCADORIAS

## UM ESTUDO METODOLÓGICO

Inês Rodrigues (ines.rodrigues@ine.pt)

6 abril 2018

# Sumário

- O que é o *Web Scraping*?
- Como é utilizado no âmbito do Inquérito ao Transporte Rodoviário de Mercadorias?
- Quais as vantagens e desafios da sua utilização?
- Como podemos contornar estes desafios?
- Considerações finais

# Web Scraping

- Técnica computacional que usa a Web como fonte de informação

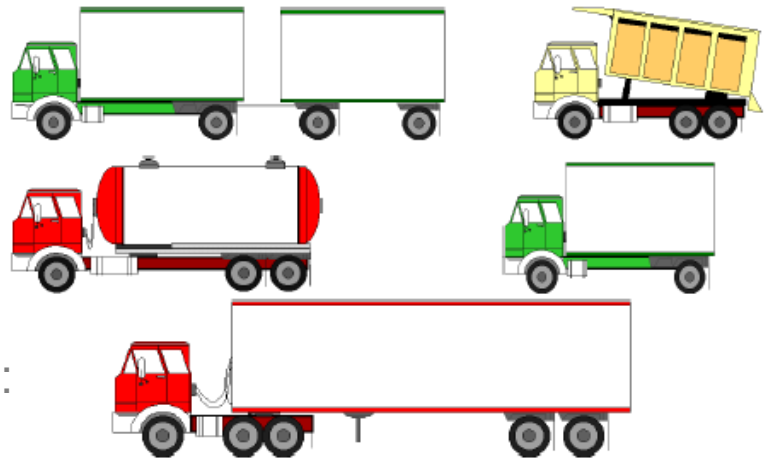

Fonte: Fernandes, M.J. (submetido) O Uso do Web Scraping nas Estatísticas Oficiais, CLADMAp III

**Automatiza o processo de recolha de dados através da Web**

# ITRM

## ITRM: Inquérito ao Transporte Rodoviário de Mercadorias

- Unidade estatística de observação:
  *veículo pesado de mercadorias*
- Inquérito amostral
- Periodicidade trimestral



Fonte: Eurostat (2016) Road Freight Transport Methodology, European Union

Atualização anual da Base de Amostragem:

- Instituto dos Registos e do Notariado, I.P. (IRN)
- Instituto da Mobilidade e dos Transportes, I.P. (IMT)

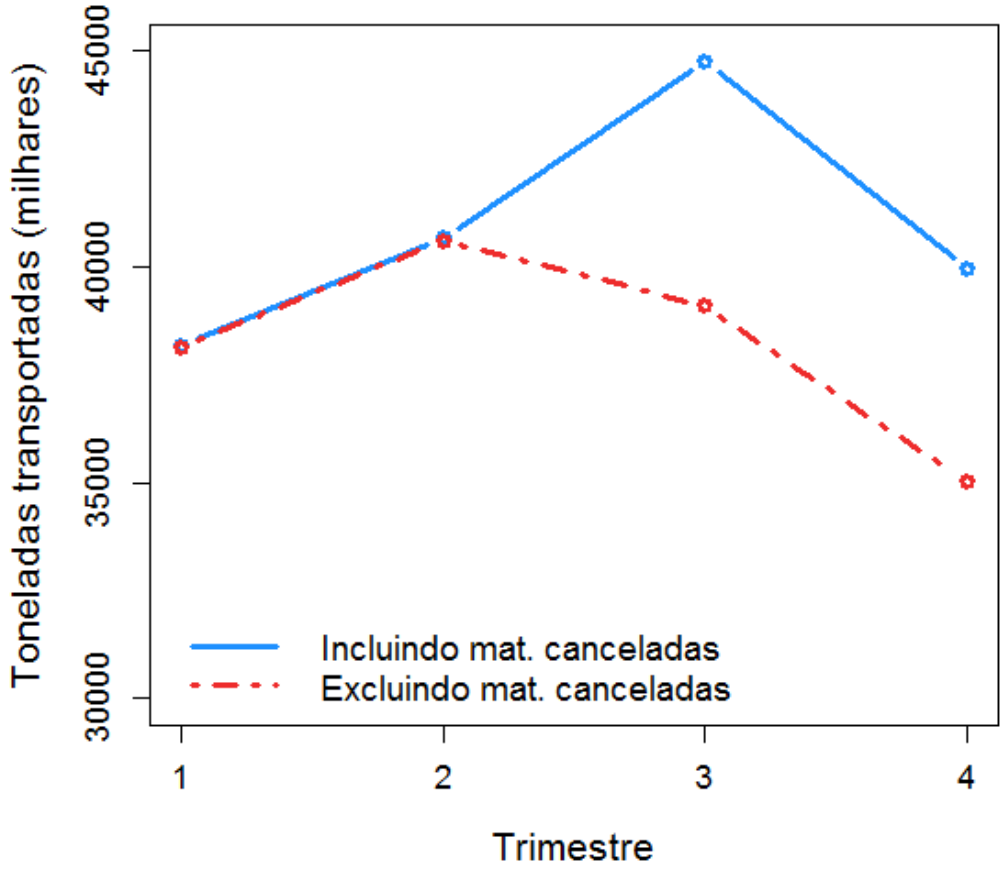
# *Web Scraping* no ITRM

# *Web Scraping* no ITRM

| Vantagens | Desafios |
| --- | --- |
| Dados fiáveis e atualizados sobre a situação das matrículas (canceladas ou não) | Ausência de informação igualmente atualizada sobre os registos de novas matriculas |
| Atualização do universo, base de amostragem e amostras trimestrais | A exclusão das matrículas canceladas da BA, ao longo do ano, conduz necessariamente à diminuição da sua dimensão |
| Melhoria nas taxas de respostas | Efeitos no cálculo dos ponderadores e estimativas trimestrais |

# *Web Scraping* no ITRM

» Toneladas transportadas, ITRM 2015 (*web scraping* a partir do 3º trimestre)



# Como podemos contornar estes desafios?

Estudar a possibilidade de implementação de um **fator de correção para a entrada de novas matrículas**, a aplicar à dimensão da base de amostragem em cada trimestre

# Objetivos do estudo

1. Estimar o número de novas matrículas entre o final de 2011 e o final de 2017;

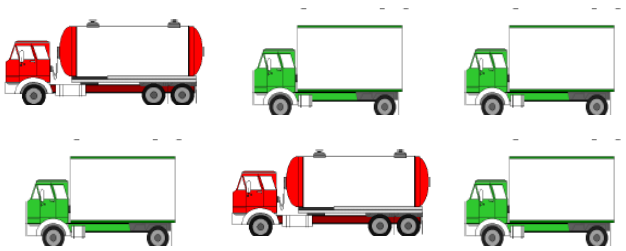
2. Analisar das diferenças entre estratos relativamente à entrada de novas matrículas;

3. Estudar modelos que permitam estimar o número anual de novas matrículas, por estrato.

# Métodos

## Nº anual de novas matrículas

Nº de matrículas presentes na base de amostragem do ano $x+1$ e que não estavam incluídas na base de amostragem do ano $x$, no estrato $h$ $(h = 1, \ldots, H)$



**Ano $x$**

**Ano $x+1$**

# Métodos

$Y_h$: nº anual de novas matrículas no estrato $h$ $(h = 1, \ldots, H)$

**Modelo de regressão de Poisson**

$$\log(\mu_h) = \eta_h = \mathrm{X}_h^T \beta$$

» $\mu_h$: valor médio condicional de $Y_h$

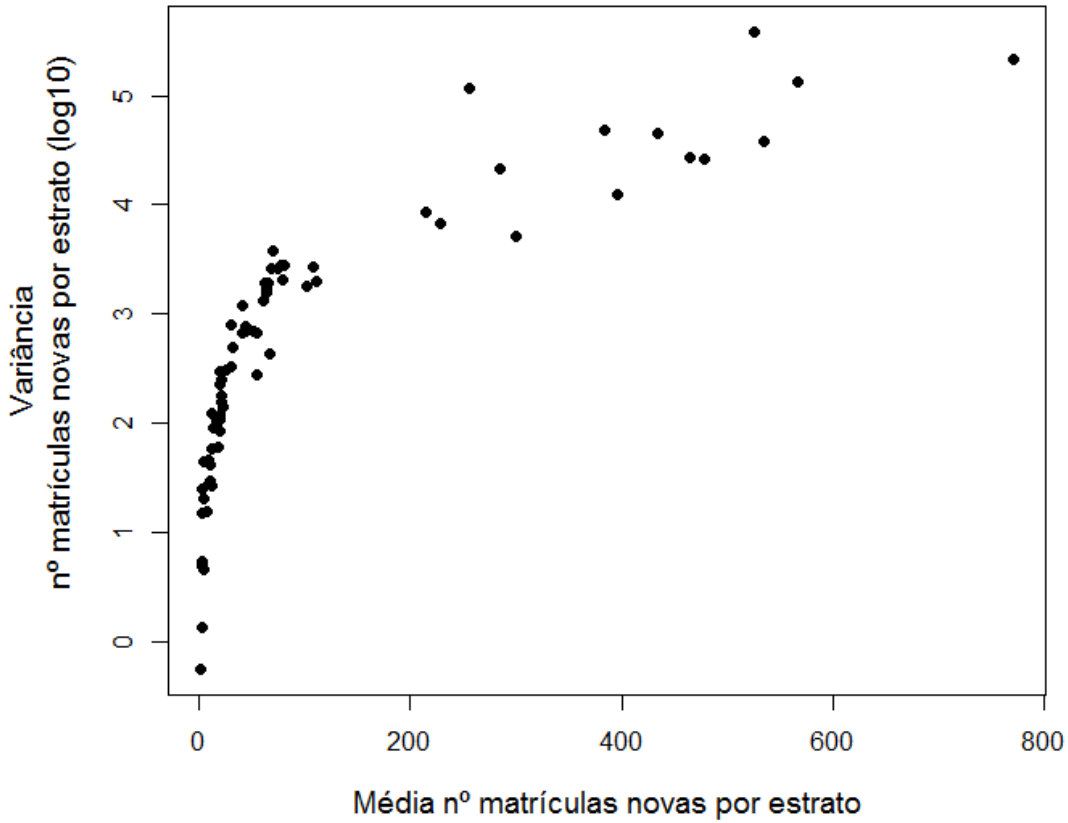
» $X_h$: vetor de covariáveis

**Constante + Variáveis de estratificação: região NUTS II, categoria do veículo, peso bruto/tara e tipo de parque**

» $\beta$: vetor de coeficientes de regressão associados a $X_h$

# Métodos

» Modelo assume que $Var(Y_h|X_h) = \mu_h$



# Métodos

## Como lidar com a sobredispersão dos dados?

### Modelo de regressão de Quasi-Poisson

$$Var(Y_h) = \phi \mu_h$$

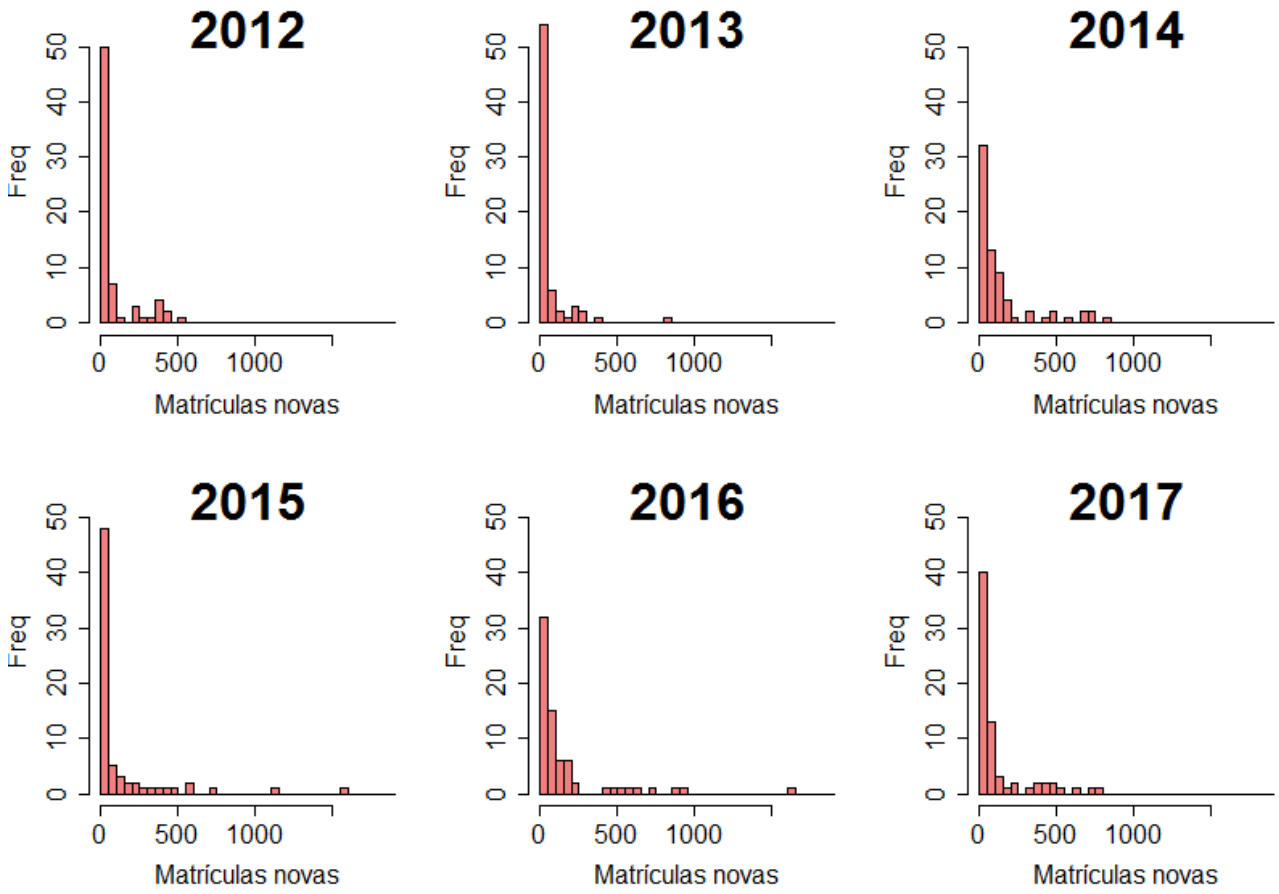
### Modelo de regressão Binomial Negativa (mistura Poisson-Gama)

$$Y_h \sim Pois(K_h) \qquad E[Y_h] = \theta/\lambda_h = \mu_h$$

$$K_h \sim \Gamma(\theta, \lambda_h) \qquad Var(Y_h) = \mu_h + \mu_h^2/\theta$$

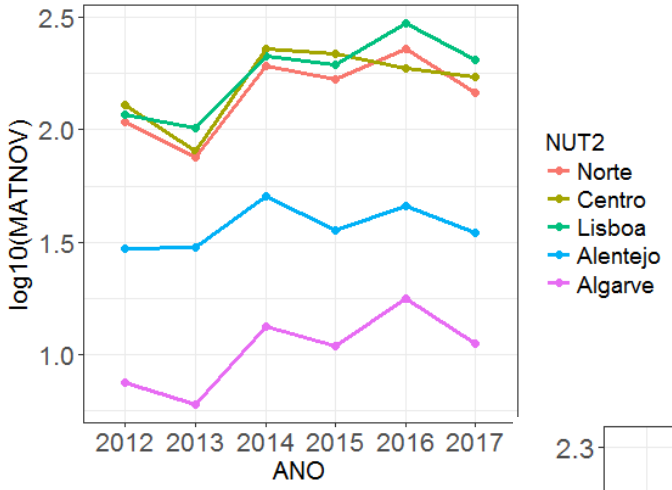
# Resultados

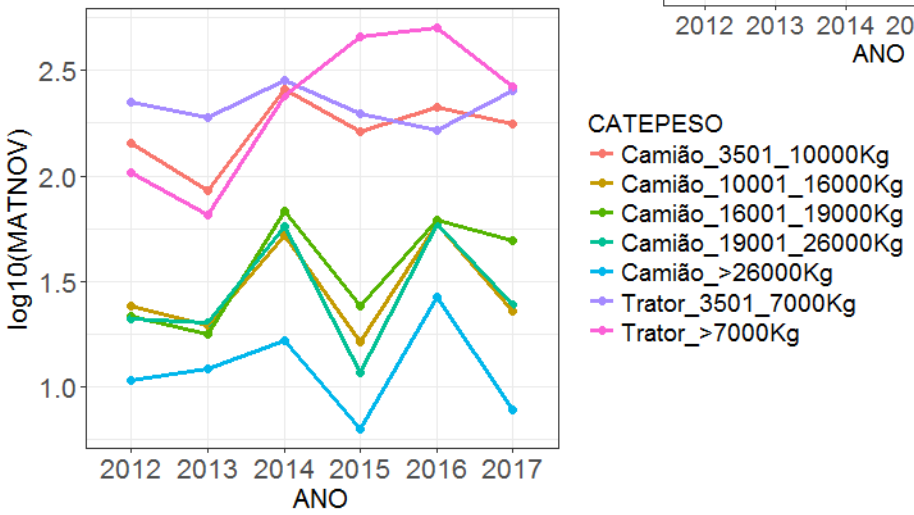
» Distribuição do número de novas matrículas, por estrato, 2012 - 2017

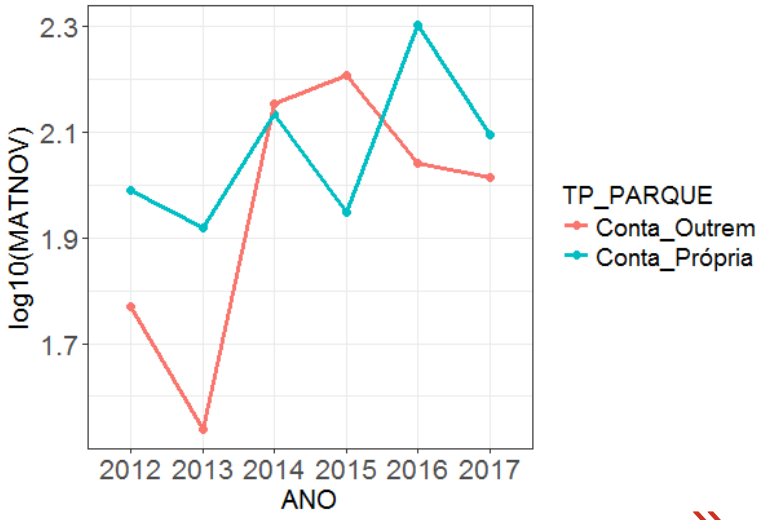


# Resultados

» Média do nº de novas matrículas por categoria de variável de estratificação, 2012 - 2017

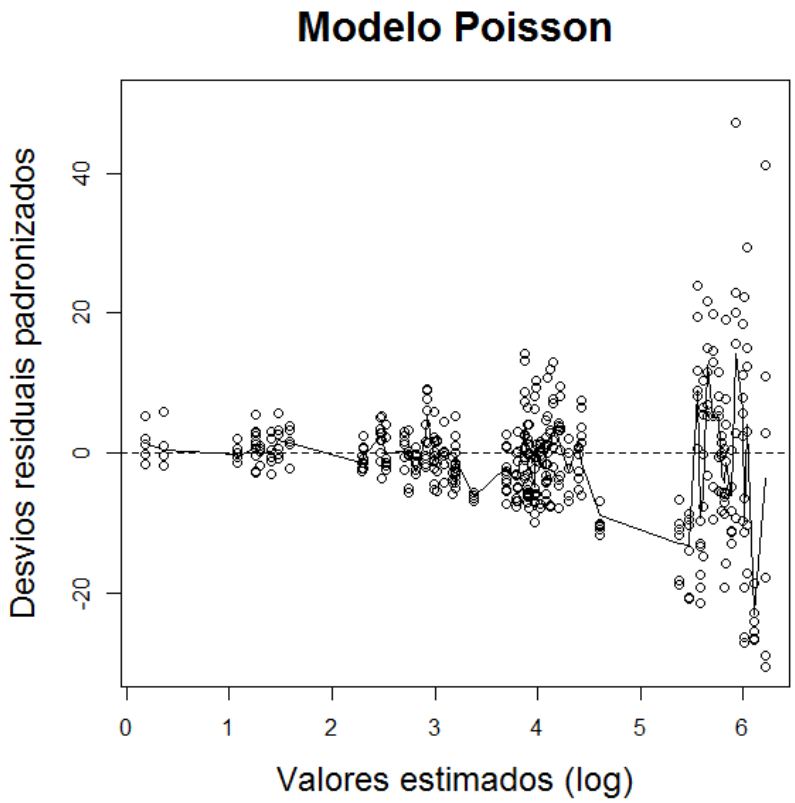


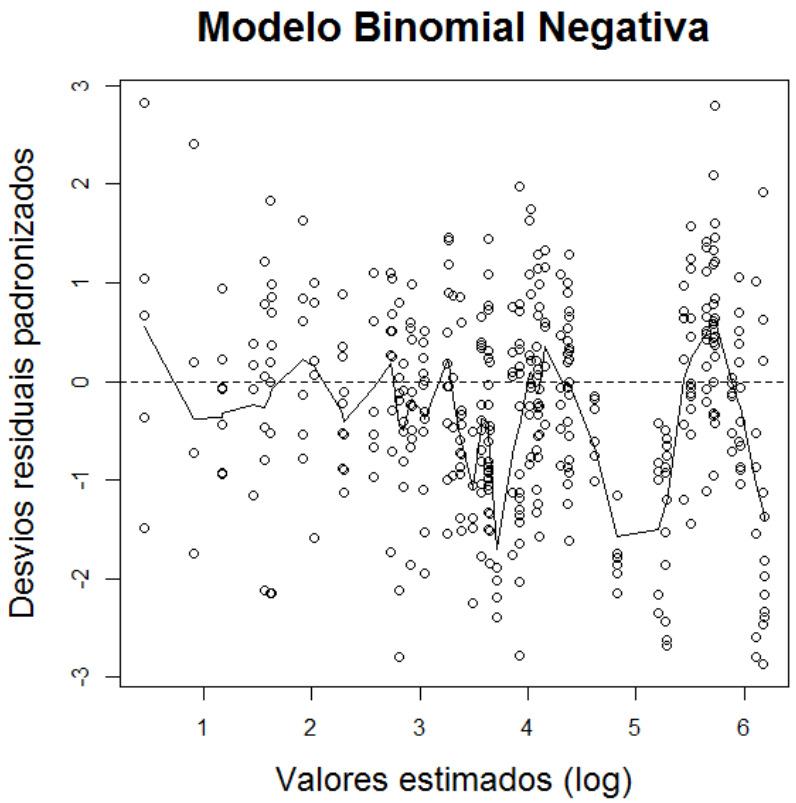



# Resultados

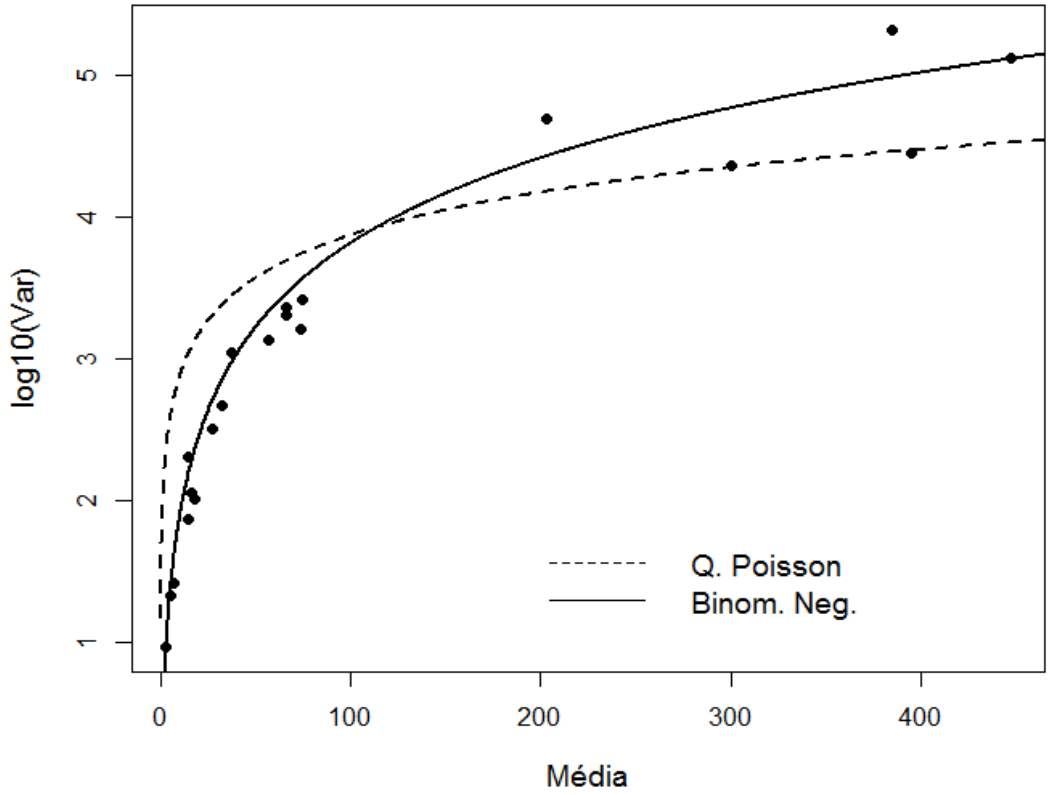
» Desvios residuais vs. valores estimados



# Resultados

» Relação média-variância estimada pelos modelos de regressão Quasi-Poisson e Binomial Negativa



# Resultados

| | 1: Bin.Negativa + Cat_Peso*Tp_Parque | 2: Bin.Negativa + Cat_Peso*Tp_Parque + Ano | 3: Bin.Negativa + Cat_Peso*Tp_Parque + Ano + Cat_Peso*Ano |
|---|---|---|---|
| | β (se) | β (se) | β (se) |
| **Constante** | 4,3 (0,15) *** | 3,88 (0,16) *** | 3,77 (0,19) *** |
| **Região (NUTS II)** | | | |
| **Norte** | ref. | ref. | ref. |
| **Centro** | -0,02 (0,11) | -0,01 (0,11) | -0.004 (0,10) |
| **AML** | 0,15 (0,11) | 0,15(0,11) | 0,13 (0,10) |
| **Alentejo** | -1,35 (0,11) *** | -1,32 (0,11) *** | -1,31 (0,10) *** |
| **Algarve** | -2.42 (0.12) *** | -2,44 (0,12) *** | -2,42 (0,11) *** |
| **Categoria de veículo e escalão de peso bruto/tara** | | | |
| **Camião** | | | |
| **3 501 - 10 000 Kg** | ref. | ref. | ref. |
| **10 001 - 16 000 Kg** | -0,96 (0,2) *** | -0,92 (0,19) *** | -0,69 (0,26) ** |
| **16 001 - 19 000 Kg** | -0,4 (0,19) * | -0,38 (0,19) * | -0,41 (0,26) |
| **19 001 - 26 000 Kg** | -0,91 (0,2) *** | -0,85 (0,19) *** | -0,61 (0,26) * |
| **Mais de 26 000 Kg** | -1,3 (0,2) *** | -1,25 (0,19) *** | -0,92 (0,27) *** |
| **Trator** | | | |
| **3 501 - 7 000 Kg** | 1,39 (0,19) *** | 1,45 (0,18) *** | 1,87 (0,25) *** |
| **Mais de 7 000 Kg** | 1,85 (0,19) *** | 1,88 (0,18) *** | 1,32 (0,25) *** |
| **Tipo de Parque** | | | |
| **Por conta própria** | 1,85 (0,19) *** | 1,93 (0,18) *** | 1,96 (0,17) *** |
| **Por conta de outrem** | ref. | ref. | ref. |



· *p-value* < 0,1, * *p-value* < 0,05, ** *p-value* < 0,01, *** *p-value* < 0,001

# Resultados

| | 1: Bin.Negativa + Cat_Peso*Tp_Parque | 2: Bin.Negativa + Cat_Peso*Tp_Parque + Ano | 3: Bin.Negativa + Cat_Peso*Tp_Parque + Ano + Cat_Peso*Ano |
|---|---|---|---|
| | β (se) | β (se) | β (se) |
| **Categoria/Peso x Tp.Parque** | | | |
| **Camião E2 - C. própria** | -0,95 (0,27) *** | -0,98 (0,26) *** | -1,03 (0,25) *** |
| **Camião E3 - C. própria** | -1,44 (0,27) *** | -1,5 (0,26) *** | -1,52 (0,25) *** |
| **Camião E4 - C. própria** | -1,03 (0,27) *** | -1,1 (0,26) *** | -1,13 (0,25) *** |
| **Camião E5 - C. própria** | -1,59 (0,28) *** | -1,7 (0,27) *** | -1,71 (0,25) *** |
| **Trator E1 - C. própria** | -1,92 (0,26) *** | -1,94 (0,25) *** | -2,01 (0,24) *** |
| **Trator E2 - C. própria** | -2,8 (0,26) *** | -2,97 (0,25) *** | -3,21 (0,24) *** |
| **Ano'** | | 0,14 (0,02) *** | 0,17 (0,05) *** |
| **Categoria/Peso x Ano'** | | | |
| **Camião E2 - Ano'** | | | -0,08 (0,1) |
| **Camião E3 - Ano'** | | | 0,01 (0,09) |
| **Camião E4 - Ano'** | | | -0,09 (0,1) |
| **Camião E5 - Ano'** | | | -0,13 (0,1) · |
| **Trator E1 - Ano'** | | | -0,16 (0,09) * |
| **Trator E2 - Ano'** | | | 0,24 (0,09) *** |
| **nº parâmetros** | 19 | 20 | 26 |
| **log L** | -1 970 | -1 952 | -1 936 |
| **TRV** | 1 vs. BinNeg: 104,98 *** | 2 vs. 1: 35,7 *** | 3 vs. 2: 31,6 *** |
| **θ (se)** | 1,99 (0,14) | 2,17 (0,16) | 2,35 (0,17) |
| **AIC** | 3 977,5 | 3 943,8 | 3 924,2 |



· *p-value* < 0,1, * *p-value* < 0,05, ** *p-value* < 0,01, *** *p-value* < 0,001

# Considerações finais

- Os modelos estudados permitem estimar o número anual de entradas, podendo depois assumir-se uma distribuição uniforme ao longo do ano;

- No entanto, as entradas de matrículas em cada estrato poderão estar associadas por corresponderem a veículos registados sob a mesma empresa;

- As covariáveis referem-se essencialmente a características dos veículos; estudar adição de efeito associado à empresa, ou tipo de empresa (ex: atividade económica, dimensão);

- A autocorrelação entre as contagens deverá também ser investigada.

# Referências


- T Booth, J.G., Casella, G., Friedl, H. & Hobert, J.P. (2003) Negative binomial loglinear mixed models. *Statistical Modelling*, 3, 179-191.

- Eurostat (2016) Road Freight Transport Methodology, European Union.

- Fernandes, M.J. (submetido) O Uso do Web Scraping nas Estatísticas Oficiais, CLADMAp III.

- Ver Hoef, J.M. & Boveng, P.L. (2007) Quasi-Poisson vs. Negative Binomial Regression: How should we model overdispersed count data? *Ecology*, 88(11), 2766-2772.

- Winkelmann, R. (1995) Duration dependence and dispersion in count data models. Journal of Business and Economic Statistics, 13(4), 467-474.

Obrigada pela vossa atenção!

**Inês Rodrigues (ines.rodrigues@ine.pt)**