Salaries in the industry, trade and services in 2011 in France

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística
Màster en Estadística i Investigació Operativa

Author : Maxime Jurado and Mathieu Marauri

Barcelona, Spain

Contents

Abstract

We analyzed the impact of the gender, socio-professional categories and time (part-time / full-time job) on the salary. We performed Bayesian Gaussian Regression and Gaussian hierarchical Bayesian regression (with random effects) to carry this study. Both of them leaded to same conclusion. If we order the socio-professional categories from 1 to 5, 1 being considered the “best”, people from category 1 will earn more than people in 5. We also see that men earn more than women.

Introduction

The structure of the salary in the industry, the services or the trade is a matter of importance in the society. It can reflects some inequalities, regarding the gender for instance, or some tendencies. The aim of this study is to analyse the effect of several indicators such as the Socio-Professional Category or the gender on the salary.

Performing such an analysis on data coming from France is for us really interesting because we have several preconceptions about the impact of the gender for example. Studying those data will allow us to either confirm or deny these preconceptions.

By performing Bayesian models the effects of the indicators will be known. Those effects will be quantified. It will provide useful knowledge on the structure of the salary.

Dataset

The dataset comes from the *Institut National de la Statistique et des Etudes Economiques* or INSEE in France. It is the statistical institute of France. The dataset contains 33 sectors of activity that are identified by the *id* variable. The classification can be found in table 8.

The variables reported in this study are listed below:

- Time is an indicator which takes value 1 for a full-time work and 0 for a part-time work. *time*
- Sexe is a binary variable with value 1 for male and 0 for female. *sexe*
- Socio-Professional Category: it takes values between 1 and 5. Table 7 shows the classification. *spc*
- Salary: it is the response variable. *salary*

The salary is the average gross income for an hour. For instance it can be the average income of female employees working full-time in the extraction industry.

In the original dataset there were a cell by combination of the different variables. It was built so that no cell contains less than 5 entries and that no entry would represent more than 80% of the total. Some data are missing due to statistical privacy. It is the case for the id 8 which is the sector of Transportation equipment manufacture.

Statistical methods

In this study several Bayesian models will be performed, first a Gaussian Bayesian model then a hierarchical Bayesian model. In both cases covariates will be add to an initial model and by comparing the Deviance Information Criterion or DIC. Since a model with more parameters will always be preferable to a “simpler” one, parsimony will also be used to select the best model. The goal is to have the best model but also the one that is

Gaussian Bayesian regression

In the Gaussian Bayesian model the point is to get the mean effect of the covariates on the salary. By performing such a model one wants to know how the covariates impact the salary. Among the different models that can be performed the best one will be used as the initial model in the hierarchical Bayesian models selection. The Gaussian Bayesian model is of this form:

$$Y|X\ N(\beta X, \tau_1)$$

$$\Pi(\beta)\ N(\mu, \tau_2)$$

$$\Pi(\tau_1)\ Gamma(a, b)$$

The parameter β is a vector and X is the matrix of the covariates. The statistical model Gaussian with a linear expression of the covariates and the intercept as a mean and a variance of τ_1. β also follows a Gaussian distribution. Finally τ_2 follows a Gamma with parameters a and b that are fixed to 0.01.

Gaussian hierarchical Bayesian regression

A hierarchical Bayesian model is a model with random effects. Those random effects allow the model to add variance to the estimates of the parameters of the covariates. It means that each sector has different values for the estimates of the covariates. This way each sector has more precise estimates and it is possible to see the differences between sectors.

The initial model is the Gaussian Bayesian model that was selected. Then random effects are added to this model and the best model is selected based on the same selection process than before.

The Gaussian hierarchical model is designed as follow:

$$Y|X\ N(\beta X + b_{1i} X_r + b_{2i}, \tau_1)$$

$$\Pi(\beta)\ N(\mu, \tau_2)$$

$$b_i\ N(0, \tau_{3,i})$$

$$\Pi(\tau_1)\ Gamma(a_1, b_1)$$

$$\Pi(\tau_{3,i})\ Gamma(a_2, b_2)$$

As previously X is a matrix of the covariates and X_r is also a matrix of the covariates but it may be different from X since random effects can be added on different covariates. Again as before $a_1, b_1, a_2 and b_2$ are all equal to 0.01.

Descriptive analysis

An exploration of the data through a descriptive analysis gives first insights on the behavior of the variables. The point here is to see graphically if some variables seem to have an impact on the response variable, the salary. A basic linear regression and some tests give also some insights.

Influence of the variable *spc*

The first idea is that the salary is supposed to be different based on the different Socio-Professional Categories. The following graph (Figure 1) shows the mean of the salary by SPC.

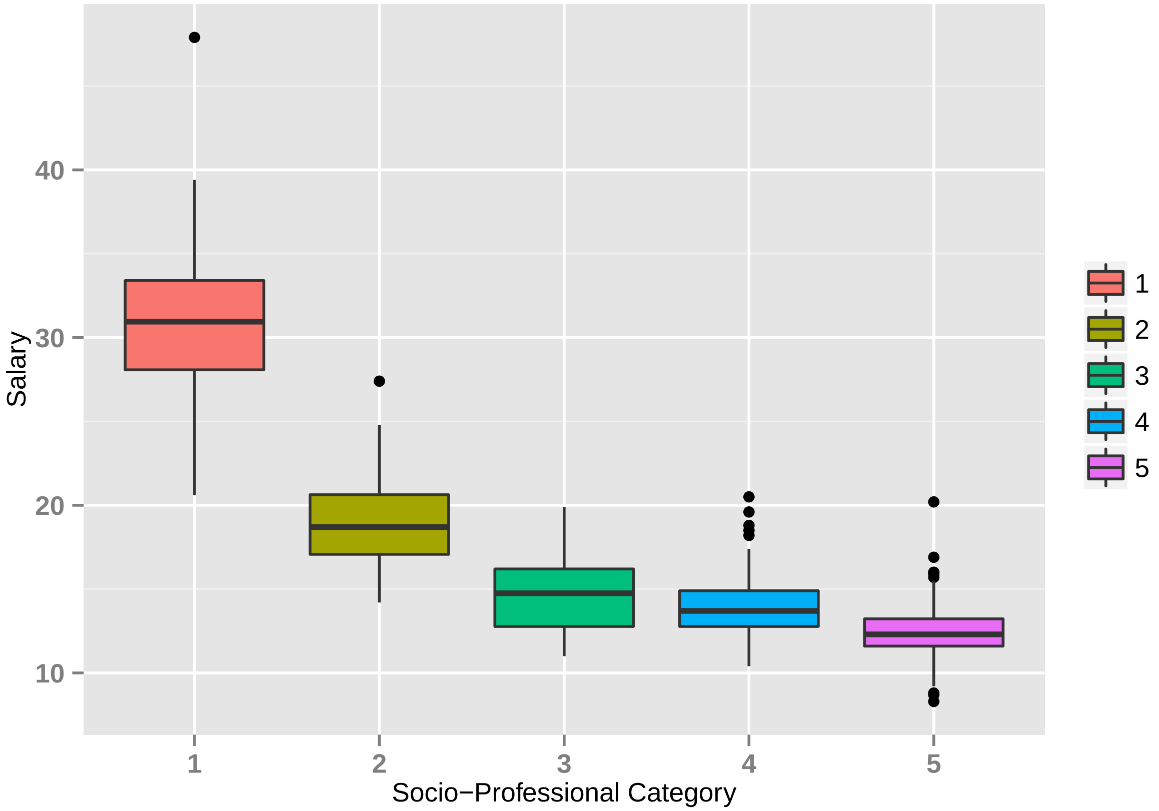

Figure 1: Boxplot of the salary for each category.

One can clearly see that being in the category 1 (Higher managerial and professional positions) increases the salary as compared to the other categories. If one considers the categories to be ordered from 1 to 5 then he would be able to say that *SPC* seems to have a negative impact on the salary.

In order to see more clearly the way *SPC* influences the salary Figure 2 shows the evolution of the salary for each SPC for each sector. The means were calculated on all the values of salary for the SPC by id.

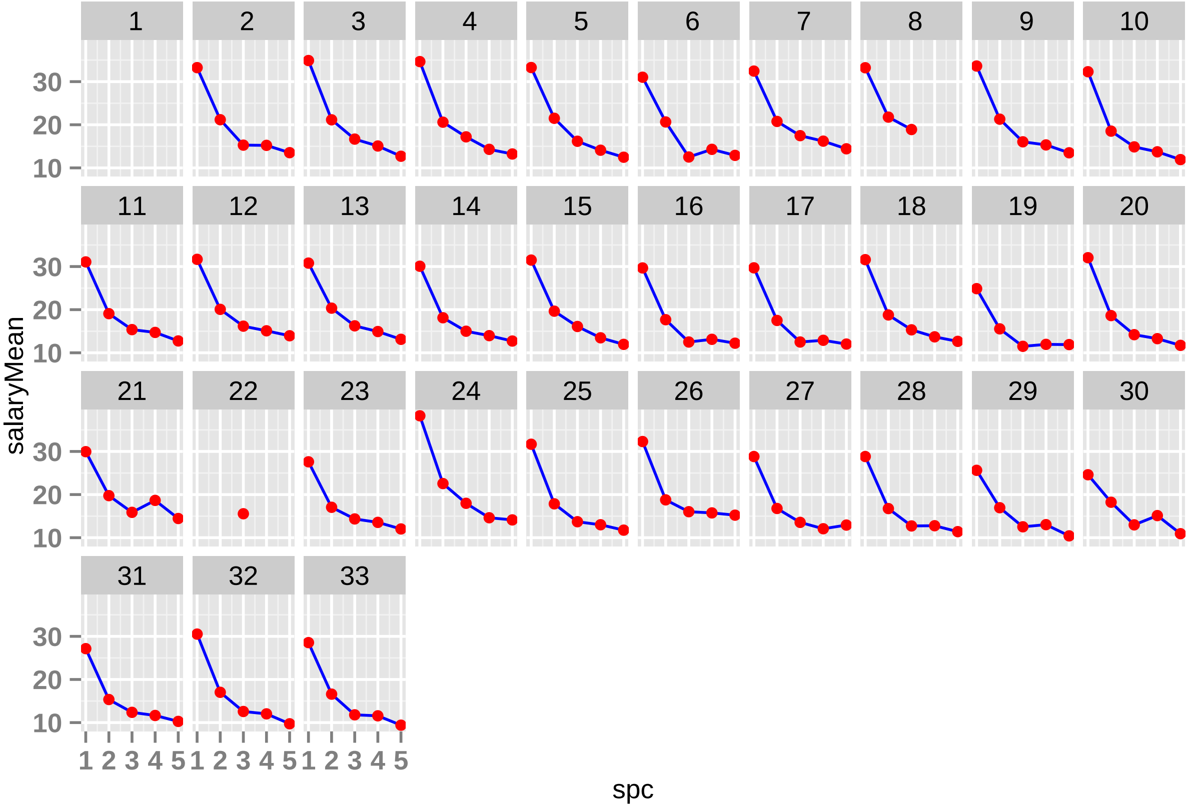

Figure 2: Evolution of the salary for each SPC by id.

As seen previously *SPC* seems to have a negative impact on the salary for every sectors but sector 6 (Manufacture of food, drink and tobacco based products), sector 21 (Information and communication) and sector 30 (Arts and entertainment). This may be due to the definitions of the SPC. In most of the sectors they are ordered from 1 to 5 but in some other sectors this order can be different.

One can see that the relation between *spc* and *salary* seems to be quadratic. Therefore it could be useful to add a quadratic part to the *spc* variable in the next regression models.

Influence of the variable *time*

The salary is given per hour, hence no differences between full-time and part-time jobs should be observed. The following figure confirms this.

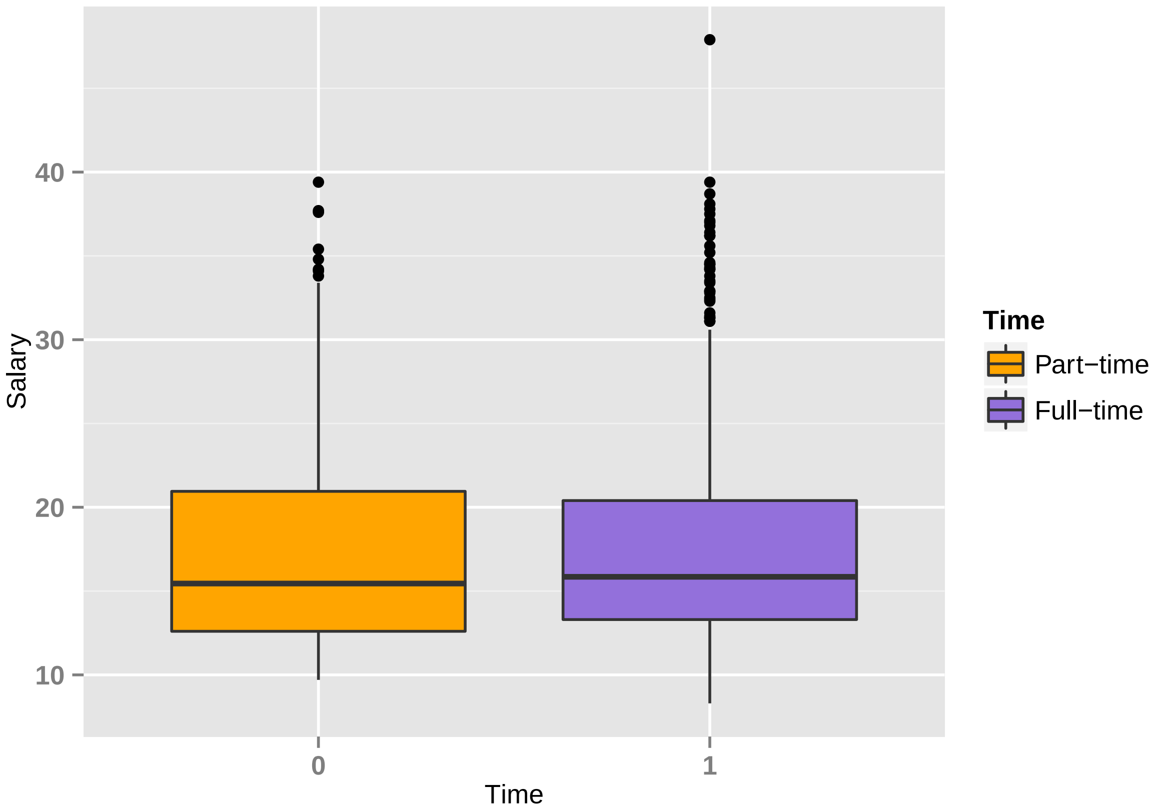

Figure 3: Boxplot of the salary for each time.

The salaries for the full-time workers ant the salaries for the part-time workers seem to be the same.

The same graph was repeated for each category and the following results were obtained. Figure 4.

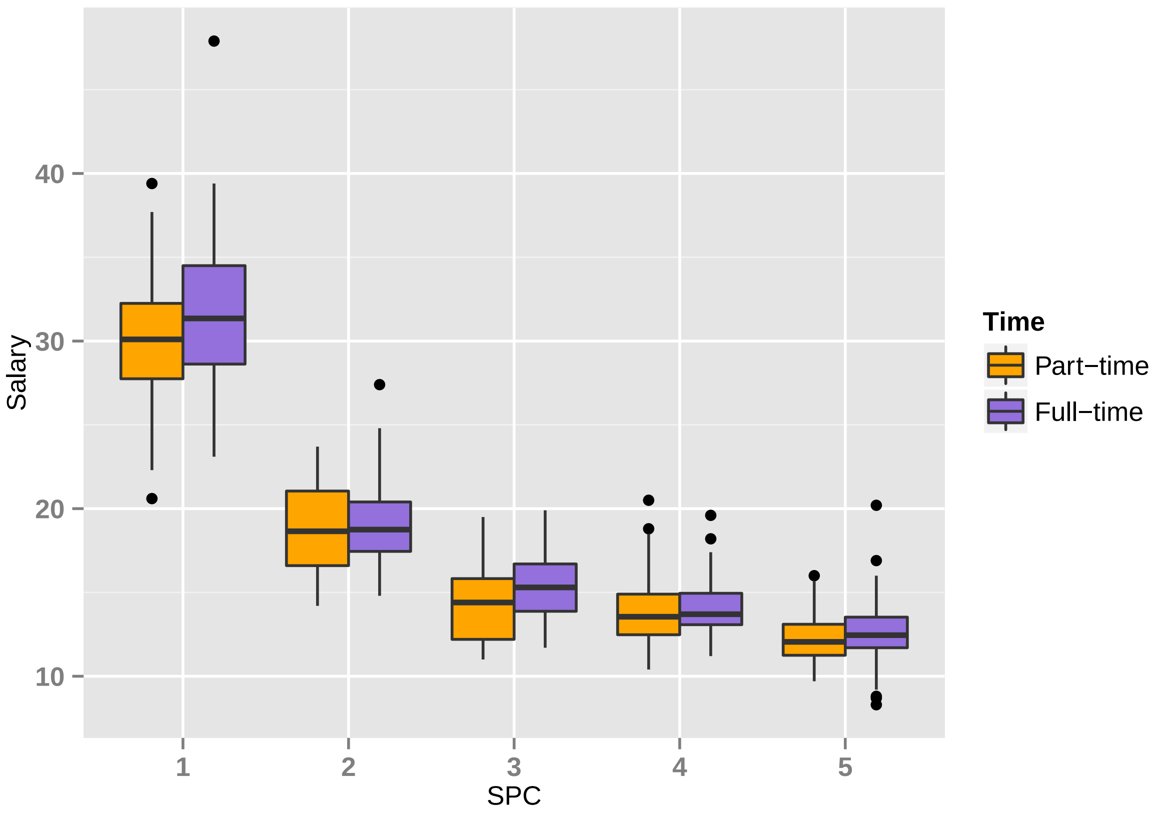

Figure 4: Boxplot of the salary for each SPC and each time.

One could say that for some categories a small difference is observed. For instance salaries for full-time workers in the category of the higher managerial and professional positions seem to be higher than the ones for the part-time workers.

A t-test was performed to see if there is a difference between the salaries of full-time workers and part-time workers. This test was performed on the whole dataset and then by category. Table 1 shows the results.

	p-value	mean of the differences
dataset	0.000	-0.788
category 1	0.000	-1.982
category 2	0.298	-0.277
category 3	0.000	-1.027
category 4	0.316	-0.188
category 5	0.051	-0.427

Table 1: P-values and mean of the differences of the tests.

One can see that the p-values for the tests performed on the whole dataset, on the category 1 and on the category 3 are below 0.05. In those cases the mean of the differences can be up to almost 2 euros per hour. It represents a difference of 6.2%.

Influence of the variable *sexe*

Inequality in the salary regarding the gender is a matter of discussion in France. The issue is that salaries are said to be higher for males than they are for females. Therefore some differences are expected.

The means of the salaries for females and males were plotted for each Socio-Professional Category and it seems that no differences appear except maybe for the category 1. Figure 5 shows the results.

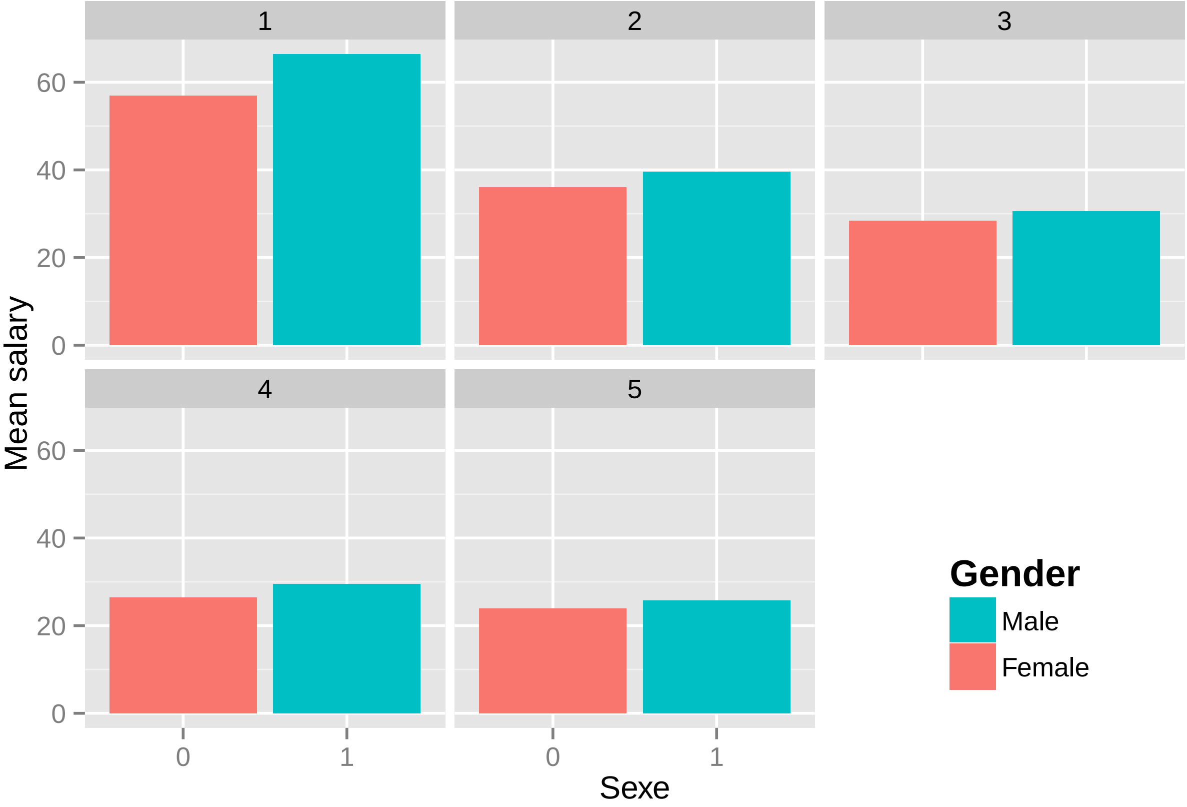

Figure 5: Barplot of the mean salary by sexe for each category.

For every categories the mean salary for males seems to be higher than the mean salary for females. A t-test confirmed this result. A significant mean of the differences of 2 was observed. It means that males have in mean a salary per hour higher by 2 euros than the one of the females.

A linear regression

A basic linear regression was performed to see the impact of the covariates on the salary. The model that was used is: $Salary = \beta_0 + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 sexe * spc + \beta_4 * spc^2$

The following table (Table 2) shows the estimates and the p-values obtained with the linear regression model.

	estimates	p-value
(Intercept)	40.218	0.000
time	0.788	0.000
sexe	4.370	0.000
spc	-14.144	0.000
spc^2	1.728	0.000
sexe:spc	-0.798	0.000

Table 2: Results of the linear regression.

As for the graphical analysis *spc* has a negative impact on the salary (we consider here that the categories are ordered from 1 to 5). The variable *sexe* has a quite important positive effect on the salary. It means that males are paid around 4 euros per hour more than females. The fact that the interaction between *sexe* and *spc* has a negative effect on the salary means that the positive effect of being a male does not counterbalance the negative impact of the *spc*. Finally *time* has a small positive impact. It means that when you work in a full-time job, you earn around 0.8 euros more than a part-time job.

Gaussian Bayesian model

Before going trough the interpretation of a Bayesian Gaussian Regression, we have to choose which model is the best. For that we start with a simple model with three covariates: *time*, *sexe* and *spc*. Then we tried to add spc^2 as explained in the descriptive part. Finally we try to add some interaction between covariates. To decide which model is the best, we use the DIC criteria and the error from the prediction. The results are presented below

	DIC	res
$\alpha + \beta_1 time + \beta_2 sexe + \beta_3 spc$	3455.60	9686.56
$\alpha + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 spc^2$	2988.00	4518.68
$\alpha + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 spc * sexe$	3444.80	9492.39
$\alpha + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 spc * sexe + \beta_5 spc^2$	2962.90	4324.22

Table 3: Table for the selection of the Bayesian models.

As you can see, the best model is the one with spc^2 and the interaction between *spc* and *sexe*. However the sum of squares is still big. It means that our model does not predict well the future values. That is why we try to add some random effects in the next part. Now we have selected the model, we can go to the interpretation.

In table 4 we can see the results of the Bayesian Gaussian Regression that we choose above. The coefficients are very similar to those obtained in the descriptive part. Thus the interpretation is the same as in ReferencesA linear regression. We still have this negative impact of *spc* and the interaction between *sexe* and *spc*. We also find again the positive impact of *sexe* that does not counterbalance the effect of *spc*. Finally the impact of *time* is still positive.

	mean
intercept	40.0549
time	0.8032
sexe	4.4220
spc	-14.0450
spc*sexe	-0.8104
spc²	1.7135

Table 4: Mean for the Bayesian model.

Gaussian hierarchical Bayesian model

As for the Bayesian Gaussian Regression, we have to choose here the best model. We proceed the same way, trying to ass random effects to one covariate at a time. But at the end, we do not want to over-parameterize the model, that is why we limit this study to 2 random effects maximum.

	DIC2	res2
$\alpha + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 spc * sexe + \beta_5 spc^2$	2962.90	4324.22
$\alpha + \beta_1 time + (\beta_2 + b_i) sexe + \beta_3 spc + \beta_4 spc * sexe + \beta_5 spc^2$	2881.20	2279.24
$\alpha + (\beta_1 + b_i) time + \beta_2 sexe + \beta_3 spc + \beta_4 spc * sexe + \beta_5 spc^2$	2964.10	4048.85
$\alpha + \beta_1 time + \beta_2 sexe + (\beta_3 + b_i) spc + \beta_4 spc * sexe + \beta_5 spc^2$	2970.00	4208.86
$\alpha + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 spc * sexe + (\beta_5 + b_i) spc^2$	2986.10	4076.92
$\alpha + b_{1i} + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 spc * sexe + (\beta_5 + b_{2i}) spc^2$	2869.30	1901.95
$\alpha + b_{1i} + \beta_1 time + \beta_2 sexe + \beta_3 spc + \beta_4 spc * sexe + (\beta_5 + b_{2i}) spc^2$	2682.40	1104.37

Table 5: Table for the selection of the hierarchical Bayesian models.

You can see that the model with random effects on the *Intercept* and *sexe* is the best. Furthermore, the sum of squares is lower than in the Bayesian Gaussian Regression. This is our final model.

In table 6, the coefficients of our final model are presented. It can be seen that they do not differ a lot from the previous interpretations. The main part here is the random effects on *Intercept* and *sexe*. It means that we add variability for each subject. We can observe differences of each sector. But here, the variance added by the random effects does not change the effect of a covariate. The sign of its coefficient will not change. The global effect will be the same.

	mean
intercept	39.6747
time	0.6752
sexe	4.3804
spc	-13.6565
spc*sexe	-0.7962
spc^2	1.6482

Table 6: Mean for the hierarchical model.

Conclusion

The aim of this study was to describe the structure of the salary per hour in France by performing Bayesian models. After a descriptive analysis that gave us insights on the structure of the salary we performed several Bayesian regressions. The selection process led us to select a hierarchical model over a simple Bayesian model. Based on all of this it appeared that the Socio-Professional Category has a great impact on the salary; being in the first category -Higher managerial and professional positions- insures a better salary. The second major effect that we observed is the one of the gender. Males have higher salary than women. Those results were observed both in Gaussian Bayesian regressions and in hierarchical Gaussian Bayesian regressions.

As a further analysis one could try to do a cluster analysis in order to regroup the sectors that are similar in the way the salary is structured. Some data could be add to the dataset in order to specify the structure of the salary even more. For instance information on the age and on the number of workers by cells.

Bibliography

- INSEE T101 Salaire brut horaire, par âge et catégorie socioprofessionnelle simplifiée
- Introduction fo WinBUGS
- WinBUGS Training

Appendix

Classification tables

Here are presented the classification tables for the Socio-Professional Categories (Table 7) and for the Sectors (Table 8).

	Category
1	Higher managerial and professional positions
2	Intermediate occupations
3	Employee
4	Skilled worker
5	Unskilled worker

Table 7: Classification table of the Socio-Professional Categories.

	Sector
1	Agriculture, silviculture and fishing
2	Manufacturing industry, extraction industry
3	Extraction industry, energy, water, trash management and remediation
4	Extraction industry
5	Water production and distribution, remediation, trash management
6	Manufacture of food, drink and tobacco based products
7	Electric, electronic and computer components manufacture
8	Transportation equipment manufacture
9	Other industrial products manufacture
10	Textile, clothing industry, leather and shoe industry
11	Wood working, paper and printing industry
12	Manufacture of rubber and plastic materials and other non metallic products
13	Metalworking industry and production of metallic products except machine and equipment
14	Other manufacture industry, repair and installation of machines and equipment
15	Construction
16	Wholesale trade and retailing, transport, accommodation and restoration
17	Trade, automobile and motorcycle repairs
18	Transportation and stocking
19	Accommodation and food service industry
20	Diverse services
21	Information and communication
22	Publishing, audiovisual media and airing
23	Computer activities and information services
24	Financial and insurance activities
25	Scientific and technical activities, administrative and support services
26	Legal, accounting, management, architecture, control and technical analysis activities
27	Other specialized activities, scientific and technical
28	Administrative services and support activities
29	Other service activities
30	Arts and entertainment
31	Other service activities
32	Public administration, teaching, health and social action
33	Medico-social accommodation, social, and social action without accommodation

Table 8: Classification table of the Sectors.

Return to section Dataset.

Code

WinBUGS code

```
model{
for(i in 1:n){
salary[i]~dnorm(mu[i],tau)
mu[i] <- alpha + beta1*time[i] + beta2*sexe[i] + beta3*spc[i] + beta4*spc[i]*sexe[i] +
beta5*spc2[i]
}
tau~dgamma(0.01,0.01)
alpha~dnorm(0,0.01)
beta1~dnorm(0,0.01)
beta2~dnorm(0,0.01)
beta3~dnorm(0,0.01)
beta4~dnorm(0,0.01)
beta5~dnorm(0,0.01)
}

###

model{
for(i in 1:n){
salary[i]~dnorm(mu[i],tau)
mu[i] <- alpha + beta1*time[i] + (beta2+b[i])*sexe[i] + beta3*spc[i] + beta4*spc[i]*sexe[i] +
beta5*spc2[i] + b1[i]
b[i]~dnorm(0,tau2)
b1[i]~dnorm(0,tau3)
}
tau~dgamma(0.01,0.01)
alpha~dnorm(0,0.01)
beta1~dnorm(0,0.01)
beta2~dnorm(0,0.01)
beta3~dnorm(0,0.01)
beta4~dnorm(0,0.01)
beta5~dnorm(0,0.01)
tau2~dgamma(0.01, 0.01)
tau3~dgamma(0.01, 0.01)
}
```

R code

```
# Packages ----------------------------------------------------------------

library("ggplot2")
library("xtable")
```

```r
# Data --------------------------------------------------------------------

data <- read.csv2("Data/data.csv",header=TRUE)
data <- subset(data,time!="TP")

data1 <- data[rep(1:nrow(data),each=5),]
data2 <- data.frame(data$X1,data$X2,data$X3,data$X4,data$X5)
data3 <- data.frame(data1$id,data1$time,data1$sexe)

vec <- NULL
for(i in 1:nrow(data2)){
  vec1 <- data2[i,]
  vec <- c(vec,t(vec1))
}

vec <- as.data.frame(vec)
data4 <- data.frame(data3,vec)
cat <- c(rep(c(1,2,3,4,5),132))

data <- data.frame(data4,cat)
colnames(data) <- c("id","time","sexe","salary","spc")
attach(data)
salary<-as.numeric(salary)
sexe <- as.factor(sexe)
time <- as.factor(time)
spc <- as.factor(spc)


# Descriptive analysis ----------------------------------------------------

theme <- theme(plot.title = element_text(size=16, face="bold"), axis.title.x = element_text(
              size=15), axis.title.y = element_text(size=15), axis.text.x = element_text(
              size=15, face="bold"),axis.text.y = element_text(size=15, face="bold"))

# SPC
ggplot(data, aes(factor(spc), salary)) + geom_boxplot(aes(fill = factor(spc))) +
  xlab("Socio-Professional Category") + ylab("Salary") + guides(fill=guide_legend(title=NULL)) +
  ggtitle("Boxplot of the salary for each category") + theme

datamean <- aggregate(data.frame(salaryMean = data$salary), by = list(id = data$id, spc = data$spc),
                      mean)

ggplot(datamean, aes(spc, salaryMean)) + facet_wrap(facets=~ id, ncol=10) +
  geom_line(colour="blue") + geom_point(colour="red") + theme

# Time
ggplot(data, aes(factor(time), salary)) + geom_boxplot(aes(fill = factor(time))) +
  labs(x="Time", y="Salary", title="Boxplot of the salary for each time") + theme +
  scale_fill_manual(name="Time", values=c("orange", "mediumpurple"), labels=c("0"="Part-time",
  "1"="Full-time"))

ggplot(data, aes(factor(spc), salary)) + geom_boxplot(aes(fill = factor(time))) +
  labs(x="SPC", y="Salary", title="Boxplot of the salary for each SPC and each time") +
  theme + scale_fill_manual(name="Time", values=c("orange", "mediumpurple"), labels=c("0"="Part-time",
  "1"="Full-time"))

## Test on the differences between time1 and time0.
data.sort <- data[with(data, order(id, spc)), ]
row.names(data.sort) <- c(1:nrow(data.sort))
data.sort.time0 <- NULL
data.sort.time1 <- NULL
for(i in 1:nrow(data.sort)){
  newline <- data.sort[i,]
  if(i%%2==0){
    data.sort.time0 <- rbind(data.sort.time0, newline)
  }
  else{
    data.sort.time1 <- rbind(data.sort.time1, newline)
  }
}
data.time.paired <- as.data.frame(cbind(data.sort.time0$spc,data.sort.time0[,4],data.sort.time1[,4]))
colnames(data.time.paired) <- c("spc", "salary0", "salary1")

test.time.all <- t.test(data.time.paired[,2], data.time.paired[,3], paired=TRUE)

data.time.paired.spc1 <- subset(data.time.paired, spc==1)
data.time.paired.spc2 <- subset(data.time.paired, spc==2)
data.time.paired.spc3 <- subset(data.time.paired, spc==3)
data.time.paired.spc4 <- subset(data.time.paired, spc==4)
data.time.paired.spc5 <- subset(data.time.paired, spc==5)

test.time.spc1 <- t.test(data.time.paired.spc1[,2], data.time.paired.spc1[,3], paired=TRUE)
test.time.spc2 <- t.test(data.time.paired.spc2[,2], data.time.paired.spc2[,3], paired=TRUE)
test.time.spc3 <- t.test(data.time.paired.spc3[,2], data.time.paired.spc3[,3], paired=TRUE)
test.time.spc4 <- t.test(data.time.paired.spc4[,2], data.time.paired.spc4[,3], paired=TRUE)
test.time.spc5 <- t.test(data.time.paired.spc5[,2], data.time.paired.spc5[,3], paired=TRUE)

table.test.time <- data.frame(c("dataset","category 1","category 2","category 3","category 4",
                   "category 5"),c(test.time.all$p.value,test.time.spc1$p.value,
                   test.time.spc2$p.value,test.time.spc3$p.value,test.time.spc4$p.value,
                   test.time.spc5$p.value), c(test.time.all$estimate,test.time.spc1$estimate,
                   test.time.spc2$estimate,test.time.spc3$estimate,test.time.spc4$estimate,
                   test.time.spc5$estimate))
colnames(table.test.time) <- c("","p-value","mean of the differences")
print(xtable(table.test.time, align=c("c","c","c","c"), caption="P-values and mean of the differences
             of the tests. \\label{tabletesttime}", digits=3))

salary.mean.spc <- aggregate(data.frame(salaryMean=data$salary),by=list(time=data$time,
                             spc=data$spc),mean,na.rm=TRUE)
(table.test.time[2,3]/salary.mean.spc[2,3])*100

# Sexe
data.m <- aggregate(data.frame(salary.m = data$salary), by = list(sexe = data$sexe,
                    spc = data$spc,time = data$time), mean, na.rm=TRUE)

ggplot(subset(data.m,time==0),aes(x=factor(sexe),y=salary.m,fill=factor(sexe))) +
  geom_bar(stat = "identity") + facet_wrap(~ spc) +
  labs(title="Mean salary by gender and spc at time=0 (Part-Time)", x="Sexe", y="Mean salary") +
  scale_fill_discrete(name="Gender",  labels=c("Female", "Male"),guide = guide_legend(reverse=TRUE)) +
  theme(plot.title = element_text(size=16, face="bold"), legend.title = element_text(colour="black",
  size=18, face="bold"),legend.position=c(.85,.15),legend.background = element_rect(size=25),
  legend.text = element_text(size = 16))

ggplot(subset(data.m,time==0),aes(x=factor(sexe),y=salary.m,fill=factor(sexe))) +
  geom_bar(stat = "identity") + facet_wrap(~ spc) +
  labs(title="Mean salary by gender and spc at time=1 (Full Time)", x="Sexe", y="Mean salary") +
  scale_fill_discrete(name="Gender",  labels=c("Female", "Male"),guide = guide_legend(reverse=TRUE)) +
  theme(plot.title = element_text(size=16, face="bold"),legend.title = element_text(colour="black",
  size=18, face="bold"),legend.position=c(.85,.15),legend.background = element_rect(size=25),
  legend.text = element_text(size = 16))

ggplot(data.m,aes(x=factor(sexe),y=salary.m,fill=factor(sexe))) + geom_bar(stat = "identity") +
  facet_wrap(~ spc) + labs(title="Mean salary by gender and spc", x="Sexe", y="Mean salary") +
  scale_fill_discrete(name="Gender",labels=c("Female","Male"),guide = guide_legend(reverse=TRUE)) +
  theme(plot.title = element_text(size=16, face="bold"),legend.title = element_text(colour="black",
  size=18, face="bold"),legend.position=c(.85,.15),legend.background = element_rect(size=25),
  legend.text = element_text(size = 16))

## Test on the differences between time1 and time0.
data.sort2 <- data[with(data, order(id, spc, time)), ]
row.names(data.sort2) <- c(1:nrow(data.sort2))
data.sort.sexe0 <- NULL
data.sort.sexe1 <- NULL
for(i in 1:nrow(data.sort2)){
  newline <- data.sort2[i,]
  if(i%%2==0){
    data.sort.sexe0 <- rbind(data.sort.sexe0, newline)
  }
  else{
    data.sort.sexe1 <- rbind(data.sort.sexe1, newline)
  }
}
data.sexe.paired <- as.data.frame(cbind(data.sort.sexe0$spc,data.sort.sexe0[,4],data.sort.sexe1[,4]))
colnames(data.sexe.paired) <- c("spc", "salary0", "salary1")

test.sexe.all <- t.test(data.sexe.paired[,2], data.sexe.paired[,3], paired=TRUE)

# Linear regression
data.spc2 <- cbind(data,data$spc^2)
colnames(data.spc2) <- c("id","time","sexe","salary","spc","spc2")

reg <- lm(salary~time+sexe*spc, data=data)
summary(reg)

reg.spc2 <- lm(salary~time+sexe*spc+spc2, data=data.spc2)
summary(reg.spc2)

anova(reg,reg.spc2)
```

```
table.reg <- data.frame(coefficients(reg.spc2),summary(reg.spc2)[[4]][,4])
colnames(table.reg) <- c("estimates","p-value")


# Missings ---------------------------------------------------------------

dataMissing <- data[which(is.na(data$salary)),]

length(dataMissing)

id1 <- data[which(data$id==1),]
# Missing in the data
id8 <- data[which(data$id==8),]
# Secret data
id22 <- data[which(data$id==22),]
# Missing in the data


# Bayesian model ---------------------------------------------------------

library(R2WinBUGS)

path.bug <- "C:/Users/Mathieu/Documents/Cours/2A/Erasmus/Cours/Bayesian analysis/Bayesian-Project/modell
path.WBS <- "C:/Users/Mathieu/Documents/Logiciels/WinBuggs/WinBUGS14/"

Iter <- 1000
Burn <- 500
Chain <- 2
Thin <- 1
n <- nrow(data)

datalist <- list(salary=data$salary, time=data$time, sexe=data$sexe, spc=data$spc, n=n)
datalist2 <- list(salary=data$salary,time=data$time, sexe=data$sexe, spc=data$spc,
                  spc2=(data$spc)^2, n=n)

# sexe,time and spc
parameters10 <- c("alpha","beta1","beta2","beta3","tau","mu")
inits10 <- list(list(tau=1),list(tau=5))
model10 <- bugs(datalist,inits=inits10,parameters.to.save=parameters10,
               model=paste(path.bug,"modelWin10.bug",sep=""),bugs.directory=path.WBS,
               n.iter=(Iter*Thin+Burn),n.burnin=Burn,n.thin=Thin,n.chains=Chain, DIC=F,debug=T)
print(model10, digits=4)
# DIC=3455.6

# Sexe, time, spc and spc^2
parameters11 <- c("alpha", "beta1", "beta2", "beta3","beta4", "tau", "mu")
inits11 <- list(list(tau=1), list(tau=5))
model11 <- bugs(datalist2, inits=inits11, parameters.to.save=parameters11,
                model=paste(path.bug,"modelWin11.bug",sep=""),
                bugs.directory=path.WBS,
                n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model11, digits=4)
# DIC=2987.5
# sexe, time, spc and spc*sexe
parameters12 <- c("alpha", "beta1", "beta2", "beta3","beta4", "tau", "mu")
inits12 <- list(list(tau=1), list(tau=5))
model12 <- bugs(datalist, inits=inits12, parameters.to.save=parameters12,
                model=paste(path.bug,"modelWin12.bug",sep=""),
                bugs.directory=path.WBS,
                n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model12, digits=4)
# DIC=3444.8

# sexe, time, spc, spc*sexe and spc²
parameters13 <- c("alpha", "beta1", "beta2", "beta3","beta4","beta5", "tau", "mu")
inits13 <- list(list(tau=1), list(tau=5))
model13 <- bugs(datalist2, inits=inits13, parameters.to.save=parameters13,
             model=paste(path.bug,"modelWin13.bug",sep=""),
             bugs.directory=path.WBS,
             n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model13, digits=4)
# DIC=2962.9

results <- t(as.data.frame(c(model13$mean[1],model13$mean[2],model13$mean[3],model13$mean[4],
             model13$mean[5],model13$mean[6])))
colnames(results) <- c("mean")
rownames(results) <- c("intercept", "time", "sexe", "spc", "spc*sexe", "spc²")

DIC <- c(3455.6,2988.0,3444.8,2962.9)

pred10 <- model10$mean$mu
res10 <- pred10 - salary
s10 <- sum(res10^2,na.rm=TRUE)

pred11 <- model11$mean$mu
res11 <- pred11 - salary
s11 <- sum(res11^2,na.rm=TRUE)

pred12 <- model12$mean$mu
res12 <- pred12 - salary
s12 <- sum(res12^2,na.rm=TRUE)

pred13 <- model13$mean$mu
res13 <- pred13 - salary
s13 <- sum(res13^2,na.rm=TRUE)

res<-c(s10,s11,s12,s13)

tab1 <- data.frame(DIC,res)

print(xtable(tab1,align=c("c","c","c"),caption="Table for the selection of the Bayesian models.
             \\label{tableselectionmodel}"))


# Hierarchical Bayesian model ---------------------------------------------
# Random effects on time
parameters21 <- c("alpha", "beta1", "beta2", "beta3","beta4","beta5", "tau","tau2", "mu")
inits21 <- list(list(tau=1,tau2=1), list(tau=2,tau2=2))
model21 <- bugs(datalist2, inits=inits21, parameters.to.save=parameters21,
                model=paste(path.bug,"modelWin21.bug",sep=""),
                bugs.directory=path.WBS,
                n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model21, digits=4)
# DIC=2964.1

# Random effects on sexe
parameters22 <- c("alpha", "beta1", "beta2", "beta3","beta4","beta5", "tau","tau2", "mu")
inits22 <- list(list(tau=1,tau2=1), list(tau=2,tau2=2))
model22 <- bugs(datalist2, inits=inits22, parameters.to.save=parameters22,
                model=paste(path.bug,"modelWin22.bug",sep=""),
                bugs.directory=path.WBS,
                n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model22, digits=4)
# DIC=2881.2

# Random effects on spc
parameters23 <- c("alpha", "beta1", "beta2", "beta3","beta4","beta5", "tau","tau2", "mu")
inits23 <- list(list(tau=1,tau2=1), list(tau=2,tau2=2))
model23 <- bugs(datalist2, inits=inits23, parameters.to.save=parameters23,
                model=paste(path.bug,"modelWin23.bug",sep=""),
                bugs.directory=path.WBS,
                n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model23, digits=4)
# DIC=2970.0

# Random effects on spc^2
parameters24 <- c("alpha", "beta1", "beta2", "beta3","beta4","beta5", "tau","tau2", "mu")
inits24 <- list(list(tau=1,tau2=1), list(tau=2,tau2=2))
model24 <- bugs(datalist2, inits=inits24, parameters.to.save=parameters24,
                model=paste(path.bug,"modelWin24.bug",sep=""),
                bugs.directory=path.WBS,
                n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model24, digits=4)
# DIC=2986.1

# Random effects on spc^2 and the intercept
parameters241 <- c("alpha", "beta1", "beta2", "beta3","beta4","beta5", "tau","tau2", "tau3", "mu")
inits241 <- list(list(tau=1,tau2=1, tau3=1), list(tau=2,tau2=2, tau3=2))
model241 <- bugs(datalist2, inits=inits241, parameters.to.save=parameters241,
                 model=paste(path.bug,"modelWin241.bug",sep=""),
                 bugs.directory=path.WBS,
                 n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model241, digits=4)
# DIC=2869.3

# Random effects sexe and intercept
parameters221 <- c("alpha", "beta1", "beta2", "beta3","beta4","beta5", "tau","tau2", "tau3", "mu")
inits221 <- list(list(tau=1,tau2=1, tau3=1), list(tau=2,tau2=2, tau3=2))
```

```
model221 <- bugs(datalist2, inits=inits221, parameters.to.save=parameters221,
                model=paste(path.bug,"modelWin221.bug",sep=""),
                bugs.directory=path.WBS,
                n.iter=(Iter*Thin+Burn),n.burnin=Burn, n.thin=Thin, n.chains=Chain, DIC=T, debug=T)
print(model221, digits=4)
# DIC=2682.4

results2 <- t(as.data.frame(c(model221$mean[1],model221$mean[2],model221$mean[3],model221$mean[4],
                             model221$mean[5],model221$mean[6])))
colnames(results2) <- c("mean")
rownames(results2) <- c("intercept", "time", "sexe", "spc", "spc*sexe", "spc²")

DIC2 <- c(2962.9,2881.2,2964.1,2970.0,2986.1,2869.3,2682.4)

pred22 <- model22$mean$mu
res22 <- pred22 - salary
s22 <- sum(res22^2,na.rm=TRUE)

pred21 <- model21$mean$mu
res21 <- pred21 - salary
s21 <- sum(res21^2,na.rm=TRUE)

pred23 <- model23$mean$mu
res23 <- pred23 - salary
s23 <- sum(res23^2,na.rm=TRUE)

pred24 <- model24$mean$mu
res24 <- pred24 - salary
s24 <- sum(res24^2,na.rm=TRUE)

pred241 <- model241$mean$mu
res241 <- pred241 - salary
s241 <- sum(res241^2,na.rm=TRUE)

pred221 <- model221$mean$mu
res221 <- pred221 - salary
s221 <- sum(res221^2,na.rm=TRUE)

res2 <- c(s13,s22,s21,s23,s24,s241,s221)

tab2 <- data.frame(DIC2,res2)

print(xtable(tab2,align=c("c","c","c"),caption="Table for the selection of the hierarchical
Bayesian models. \\label{tableselectionmodel}"))
```