AGENIS-NEVERS Marc
Publication : 13/07/2016
Version :1.1

Différence de compétitivité entre pays du monde, analyse statistique

Contexte

En mai 2016 l'IMD (International Institute for Management Development), institut indépendant et reconnu mondialement, basé à Lausanne (Suisse), a publié son classement annuel de la compétitivité mondiale. Il se focalise sur l'étude des entreprises et des stratégies au sein d'une base de 61 pays du monde (voir la liste en annexe).

Le but de cette note est d'utiliser le score de compétitivité de l'ensemble de ces pays pour déterminer si des différences significatives apparaissent au regard de certains facteurs géopolitiques comme la population, la superficie, l'appartenance à l'Union Européenne ou à la zone euro.

Méthodologie

Indicateur de compétitivité

La méthode utilisée pour créer l'indice de compétitivité est décrite sur le site (http://www.imd.org/wcc/research-methodology/), elle consiste à agréger des indices de diverse nature:

- *Based on analysis made by leading scholars and by our own research and experience, our methodology thus divides the national environment into four main factors:*
  - *Economic Performance*
  - *Government Efficiency*
  - *Business Efficiency*
  - *Infrastructure*
- *In turn, each of these factors is divided into 5 sub-factors which highlight every facet of the areas analyzed. Altogether, the World Competitiveness Yearbook features 20 such sub-factors.*
- *These 20 sub-factors comprise more than 340 criteria, although each sub-factor does not necessarily have the same number of criteria (for example, it takes more criteria to assess Education than to evaluate Prices).*
- *Each sub-factor, independently of the number of criteria it contains, has the same weight in the overall consolidation of results, which is 5% (20x5 =100).*
- *Criteria can be hard data, which analyzes competitiveness as it can be measured (e.g. GDP) or soft data, which analyzes competitiveness as it can be perceived (e.g. Availability of competent managers). Hard criteria represent a weight of 2/3 in the overall ranking, whereas the survey data represent a weight of 1/3.*
- *In addition, some criteria are for background information only, which means that they are not used in calculating the overall competitiveness ranking (e.g. Population under 15).*
- *Finally, aggregating the results of the 20 sub-factors makes the total consolidation, which leads to the overall ranking of the World Competitiveness Yearbook.*

Méthodologie de test

La comparaison de valeur moyenne ou médiane de deux groupes de points, d'un point de vue statistique, doit faire l'objet d'un test approprié. Il s'agit principalement de vérifier si la différence de moyenne/médiane observée est telle que le hasard seul n'aurait pu l'expliquer (ou alors dans moins

de 5% des cas, ce seuil étant une valeur communément utilisée en statistiques). Parmi les nombreux tests existants, nous sélectionnons le test non-paramétrique de rang de Wilcoxon-Mann-Whitney, qui compare la médiane de deux groupes d'une population, et ne requiert aucune hypothèse préalable. Il sera utilisé dans sa version non-appariée, au seuil de rejet de 5%, symétrique (sans a priori sur le sens de la comparaison).

L'étude d'un lien entre deux séries de valeurs numériques est réalisée au moyen d'une régression linéaire $y = ax + b$. La variable x est également testée pour sa significativité, et les hypothèses de validité du modèle sont alors vérifiées (comme par exemple la normalité des résidus).

L'analyse des données est effectuée à l'aide du logiciel R-Studio couplé à R version 3.2.3.

Description des données

Les données ont été récupérées manuellement à partir du tableau communiqué à la presse, visible ici. Les données sont complètes (pas de valeurs manquantes ou aberrantes) mais ne représentent qu'à peine un tiers des 197 pays du monde reconnus par l'ONU (http://www.statistiques-mondiales.com/onu.htm).

Les données comprennent, outre le nom de chacun des 61 pays, les variables quantitatives :

- le score et le rang obtenu en 2016,
- le rang obtenu en 2015.

Rappelons que le score est calculé par une agrégation d'indices, puis subit une transformation *min-max* pour s'afficher entre 0 et 100, il ne s'agit donc pas d'une variable « bornée » *sensu stricto* qui requerrait un soin particulier (voir par exemple cette discussion) lors de sa modélisation. La distribution des scores présente ici un aspect gaussien avec une moyenne à 73 et un écart-type de 15. Un test de normalité n'est pas rejeté au seuil de 5%.

On réalise alors une jointure avec une autre source de donnée : le tableau des populations et superficie de l'ensemble des pays du globe disponible sur Wikipédia, dans sa version du 27/06/2016.

On se propose de transformer les variables 'population' et 'superficie' pour éviter des écarts trop grands (4 et 6 ordres de grandeur, respectivement, entre les valeurs extrêmes) en passant au logarithme décimal, cela ayant pour effet de normaliser la distribution des valeurs (test de normalité de Shapiro accepté avec $p - value = 0.71$).

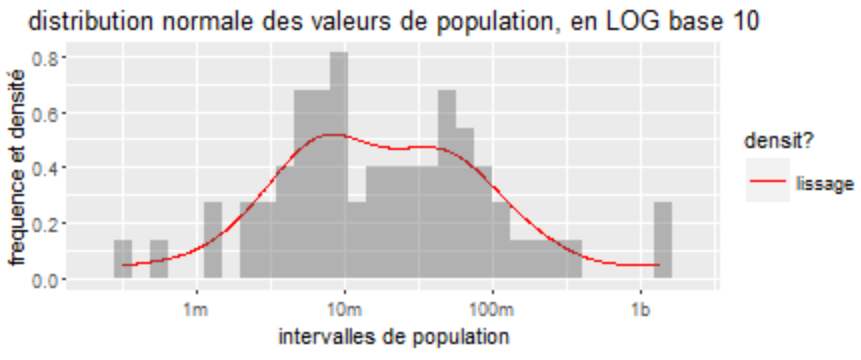

Figure 1: distribution de fréquence et densité de population dans 61 pays du monde

Test d'un lien entre superficie d'un pays et score

Le modèle linéaire (par méthode des moindres carrés) du score en fonction du logarithme de la superficie du pays montre qu'un pays comme les USA – respectivement la Chine pour le modèle population – même après transformation de la variable, continue à influencer fortement la régression, et ce n'est pas souhaitable. La présence de ces points particulièrement influents, évalués par la distance de Cook (pour tout point i, elle mesure la variation des coefficients du modèle sans ce $i^{ème}$ point), suggère de plutôt utiliser régression robuste pour l'estimation des coefficients (*estimateur M*), qui est beaucoup moins sensible aux valeurs isolées.

Tableau 1: Résultats du modèle linéaire robuste score2016~log.superficie

	Estimation	*F-stat*	*p-value*	*Validité modèle*
Intercept	91.1			OUI
Log10(superficie)	-3.51	6.17	**0.016**	

Le Tableau 1 des résultats du modèle indique clairement que **la superficie a un impact significatif** ($p - value < 5\%$) sur le score. La superficie d'un pays joue de façon inversement proportionnelle au score de compétitivité. Ainsi, **un pays 10 fois plus petit qu'un autre obtient un score meilleur de 3.5 points, en moyenne**. Le modèle donne des résultats similaires pour l'année 2015. Le R^2 du modèle, à 8%, n'est pas très élevé et confirme que de nombreux autres paramètres peuvent expliquer ce score.

Test d'un lien entre population d'un pays et score

Le même type de modèle est testé sur le logarithme de la population. Le modèle est non-significatif, c'est-à-dire qu'on ne montre **pas d'effet notable du nombre d'habitants d'un pays sur le score IMD** 2016. Les résultats sont présentés dans le Tableau 2 suivant :

Tableau 2: Résultats du modèle linéaire score2016~log.population

	Estimation	*F-stat*	*p-value*	*Validité modèle*
Intercept	104.8			OUI
Log10(population)	-4.30	2.46	**0.122**	

Test d'un lien entre appartenance à l'UE/zone Euro et le score

Au sein du continent européen on souhaite comparer les pays membres de l'UE et de la zone Euro (17 pays étudiés) avec ceux membres d'aucun des deux entités (4 pays étudiés). Le test de Wilcoxon-Mann-Whitney permet de statuer sur une **absence de différence de compétitivité selon l'appartenance ou non à la zone UE+Euro** ($p.value = 0.46$). Même si la moyenne ou la médiane donnent l'avantage aux pays hors zone UE/Euro, leur faible nombre dans ce groupe (4) limite la possibilité d'atteindre la significativité.

On n'obtient pas non plus de différence due à la zone Euro ou à l'UE par rapport au reste des pays étudiés dans le monde. Notons toutefois que deux pays membres de l'UE sont manquants dans les données (Chypre, Malte).

Test d'un modèle complet

Un modèle linéaire complet est alors testé avec l'ensemble des variables citées (log.population, log.superficie, UE, Euro), ainsi que toutes les interactions possibles de niveau 2. Une sélection exhaustive de variables (méthode '*regsubsets*') ne conserve comme significatif que le modèle étudié précédemment, avec uniquement la log.superficie comme variable explicative.

Conclusion et discussion

Nous nous sommes limités à quatre facteurs assez généraux pour l'analyse des données (population, superficie, zone Euro, zone UE). La répétition de ce genre de tests sur un plus grand nombre de facteurs explicatifs, potentiellement intéressants, devrait se faire en application de la correction de *Bonferronni* (abaissement de la $p.value$ face au risque de faux positif).

Nous concluons d'une part que nous ne pouvons pas affirmer que l'appartenance à l'UE ou l'Euro soient liées positivement ou négativement au score de compétitivité des entreprises établi par l'IMD en 2016 ou 2015. Nous concluons d'autre part en contradiction avec l'idée que pour avoir un environnement économique compétitif et dynamique il faudrait être un gros pays, en taille ou en population.

Rappelons néanmoins que le jeu de données n'est pas exhaustif des pays du monde et que la présence d'un biais ne doit pas être exclue. La donnée manquante n'est en effet pas répartie aléatoirement : de nombreux pays absents sont justement de très petite taille et certains continents comme l'Afrique sont sous-représentés.

Le tableau de données et le script R associé est à disposition du lecteur en annexe.

Annexes

Bibliographie

Les références bibliographiques sont accessibles sur le web, les liens étant fournis au fil du texte.

http://www.imd.org/wcc/research-methodology/

http://www.imd.org/uupload/imd.website/wcc/scoreboard.pdf

http://www.statistiques-mondiales.com/onu.htm

https://simple.wikipedia.org/wiki/List_of_countries_by_population_density

Tableau de données

Le tableau de données avec les variables principales, les trois variables UE, Euro et Continent désignent par 1 l'appartenance et 0 la non-appartenance à l'UE, à la zone Euro ou au continent Européen.

Pays	*rang2016*	*rang2015*	*Score2016*	*UE?*	*Euro?*	*Continent?*	*Population*	*Superficie*
Hong Kong (China)	1	2	100	0	0	0	7264100	1104
switzerland	2	4	98,018	0	0	1	7761800	41293
united states	3	1	97,881	0	0	0	321830000	9826675
singapore	4	3	97,649	0	0	0	5076700	710,2
sweden	5	9	92,353	1	0	1	9366092	449964
denmark	6	8	91,756	1	0	1	5532531	43094
ireland	7	16	91,54	1	1	1	4581269	70273
netherlands	8	15	91,321	1	1	1	17110000	41526
norway	9	7	90,054	0	0	1	5195580	385155
canada	10	5	90,048	0	0	0	33740000	9984670
luxembourg	11	6	90,016	1	1	1	502207	2586
germany	12	10	88,569	1	1	1	81757600	357022
qatar	13	13	86,716	0	0	0	1409000	11
Republic of China (Taiwan)	14	11	86,374	0	0	0	23069345	3598
United Arab Emirates	15	12	86,065	0	0	0	8264070	836
new zealand	16	17	85,606	0	0	0	4315800	270534
australia	17	18	84,27	0	0	0	23839595	7682300
united kingdom	18	19	83,338	1	0	1	62041708	24361
malaysia	19	14	83,048	0	0	0	28306700	329847
finland	20	20	82,037	1	1	1	5502075	338145
israel	21	21	80,827	0	0	0	7697600	2077
belgium	22	23	80,688	1	1	1	10827519	30528
iceland	23	24	80,58	0	0	1	318452	103
austria	24	26	80,159	1	1	1	8372930	83858
china	25	22	79,351	0	0	0	1378890000	9640821

japan	26	27	78,716	0	0	0	127387000	377873
czech republic	27	29	76,145	1	0	1	10532770	78866
thailand	28	30	74,681	0	0	0	64232760	513115
south korea	29	25	74,195	0	0	0	48456369	99538
lithuania	30	28	74,039	1	1	1	3053800	653
estonia	31	31	73,548	1	1	1	1315819	451
France (Metropolitan)	32	32	73,464	1	1	1	62793432	5515
poland	33	33	71,303	1	0	1	38163895	312685
spain	34	37	69,354	1	1	1	46087170	50603
italy	35	38	68,705	1	1	1	60200060	301318
chile	36	35	67,442	0	0	0	18086199	756096
latvia	37	43	66,554	1	1	1	2248961	646
turkey	38	40	66,551	0	0	0	77804122	783562
portugal	39	36	66,406	1	1	1	10636888	92391
slovakia	40	46	65,886	1	1	1	5424057	49033
india	41	44	65,831	0	0	0	1277401883	3287240
philippines	42	41	65,54	0	0	0	92226600	300076
slovenia	43	49	64,87	1	1	1	2164976	20256
russia	44	45	63,939	0	0	0	142905208	17098242
mexico	45	39	63,235	0	0	0	107550697	1958201
hungary	46	48	62,649	1	0	1	10013628	93032
kazakhstan	47	34	62,636	0	0	0	17010000	2724900
indonesia	48	42	62,376	0	0	0	237556363	1904569
romania	49	47	62,268	1	0	1	21466174	238391
bulgaria	50	55	61,743	1	0	1	7351234	110912
colombia	51	51	58,293	0	0	0	48638072	1138914
south africa	52	53	57,797	0	0	0	50586757	1221037
jordan	53	52	56,875	0	0	0	6316000	89342
peru	54	54	56,202	0	0	0	29461933	1285216
argentina	55	59	53,748	0	0	0	40091359	2780400
greece	56	50	52,125	1	1	1	11306183	131957
brazil	57	56	51,646	0	0	0	203773375	8514877
croatia	58	58	51,589	1	0	1	4443000	56538
ukraine	59	60	46,512	0	0	1	46936000	6037
mongolia	60	57	45,79	0	0	0	2671000	1564116
venezuela	61	61	32,603	0	0	0	31524413	916445

Script complet sous R

```
##############
# ANALYSE de la COMPETITIVITE des PAYS, rappor IMD2016
##############

# lien de la page IMD de la publication
browseURL(url="http://www.imd.org/wcc/news-wcy-ranking/")

# préparation des données
setwd("...  [ mon chemin répertoire de travail ] ...")
```

```r
library(dplyr)
library(ggplot2)
library(leaps)

# Fonctions utilisées dans ce script
import = function(file, desktop=FALSE, ...) { # entre guillemets
  path <- file
  if (desktop) {
    path <- paste0("C:/Users/", Sys.getenv("USERNAME"), "/Desktop/", file)
  }
  return(read.csv2(path, sep=";", dec=",", header=TRUE, ...))
}
HistoDens = function(x, intervalles=NULL, titre="histogramme et densit? liss?e") {
  g <- qplot(x, geom='blank') +
    geom_line(aes(y = ..density.., colour = 'Empirical'), stat = 'density') +
    geom_histogram(aes(y = ..density..), alpha = 0.4, binwidth=intervalles) +
    scale_colour_manual(name = 'densit?', values="red", label="lissage") +
    xlab("intervalles de la variable") + ylab("fr?quence & densit?") + ggtitle(titre)
  return(g)
}

# Création du tableau de données
df <- import("donnees.csv")
row.names(df) <- df$pays
# Correction des variables
df$log.population <- log10(df$population)
df$log.superficie <- log10(df$superficie)

#################
# DESCRIPTION DES DONNEES

glimpse(df)
# variable réponse (score): description
summary(df$score); message("écart-type du score:"); df$score %>% sd %>% round(2)
shapiro.test(df$score)
# écart d'ordre de grandeur pour population et surface
range(df$population) %>% log10 %>% diff
range(df$superficie) %>% log10 %>% diff
# normalité?
shapiro.test(df$log.population)
HistoDens(df$log.population, titre="distribution normale des valeurs de population, en LOG
base 10")+scale_x_continuous(labels=c(0,"1m","10m", "100m", "1b")) + xlab("intervalles de
population") + ylab("frequence et densité")

#################
# ANALYSE

# Test d'un lien entre superficie d'un pays et score
fit <- lm(data=df, score ~ log.population)
# des individus trop "influents" sur la régression normale?
which(cooks.distance(fit) > 3*length(fit$coef)/nrow(df))
# régression robuste
fit <- MASS::rlm(data=df, score ~ log.superficie)
car::Anova(fit); coef(fit)
# validation (rapide ici)
shapiro.test(fit$residuals)

# Test d'un lien entre population d'un pays et score
fit <- MASS::rlm(data=df, score ~ log.population)
car::Anova(fit); coef(fit)

# Test d'un lien entre appartenance à l'UE/zone Euro et le score
wilcox.test(x = df %>% filter(continent==1, UE==0, euro==0) %>% .$score,
            y = df %>% filter(continent==1, UE==1, euro==1) %>% .$score,
            alternative="two.sided", paired=F, exact=T, correct=F, conf.int=T)
# Test d'un lien entre appartenance à l'UE dans le monde le score
wilcox.test(x = df %>% filter(UE==0) %>% .$score,
            y = df %>% filter(UE==1) %>% .$score,
            alternative="two.sided", paired=F, exact=T, correct=F, conf.int=T)
# Test d'un lien entre appartenance à l'euro dans le monde le score
wilcox.test(x = df %>% filter(euro==0) %>% .$score,
            y = df %>% filter(euro==1) %>% .$score,
            alternative="two.sided", paired=F, exact=T, correct=F, conf.int=T)

# Test d'un modèle complet et sélection de variable
fit <- regsubsets(data=df, score ~ (log.population+log.superficie+UE+euro)^2)
summary(fit)

#################
# FIN
```