Methodology

Panel Data

This study uses the panel data methodology which is a common approach to adress the CFP-CEP nexus [@Albertini2013]. Panel data analysis is considered to be one of the most efficient analytical methods for data analysis [@DimitriosAsteriou2006]. It usually contains more degrees of freedom, less collinearity among the variables, more efficiency and more sample variability than one-dimensional method (i.e.cross-sectional data and time series data) giving a more accurate inference of the parameters estimated in the model [@Hsiao2007, @HsiaoChapitrePanelData2014]. @Roberts2013 also argued that using panel data offer a partial, but by no means complete and costless, solution to the problem of omitted variables in econometric model, namely the most common causes of endogeneity in empirical corporate finance. Panel data takes the following econometric form :

$$Y_{it} = \alpha + \beta X_{it} + u_{it} \tag{1}$$

Panel data, also called longitudinal data, includes observations on $i = 1, ..., n$ cross section units (e.g. firms) over $t = 1, ..., T$ time-periods [@Hsiao2007a]. Here Y_{it} is the dependent variable, X_{it} represents a K-dimensional row vectors of dependent variables, α is the intercept, β is a K-dimensional column vectors of parameters and u_{it} is the random disturbance term of mean equals zero. The latter which can be decomposed as $u_{it} = \mu_i + \epsilon_{it}$. The first term, μ_i, represents the individual error components and do not change over time. It can be considered as the unobserved effect model. The second term, ϵ_{it}, is the idiosyncratic error which is assumed well-behaved and independent of X_{it} and μ_i.

The starting point of all panel data is to determine if μ_i is correlated with X_{it}. In case it is correlated, then μ_i is considered as the *Fixed Effect* (i.e. FE) and the initial equation 1 is now described as the equation 2. Else, μ_i is considered as the *Random Effect* (i.e. RE) and the equation 1 becomes equation 3.

$$Y_{it} = (\alpha + \mu i) + \beta X_{it} + \epsilon_{it} \tag{2}$$

$$Y_{it} = \alpha + \beta X_{it} + (\epsilon_{it} + \mu i) \tag{3}$$

Fixed (i.e. Equation 2) and Random (i.e. Equation 3) Effect Model implies that the Ordinary Least Square (i.e. OLS) estimator of β are inconsistent[1]. Indeed FE Model violates the fourth assumption of OLS, namely an error term that is uncorrelated with each explanatory variables. RE model implies that the common error component over individuals induces correlation across the composite error terms making the third assumption of OLS violated (i.e. no linear relationship among the explanatory variables) [@Croissant2008].

While OLS is not consistent to estimate panel data model, the R package *plm* provides useful estimation methods. (i) *Pooled ols estimation* ignores the panel structure of the data and apply the same cofficients to each individual [@Schmidheiny2015]. (ii) *The random effects estimation* is the feasible Generalized Least Squares (i.e. GLS) estimator. (iii) *The fixed effects estimation* transforms the original equation 1 in substracting the time averages to every variables, such as :

$$Y_{it} - \bar{Y_i} = \beta(X_{itk} - \bar{X}_{ik}) + (u_{it} - \bar{u}_i) \tag{4}$$

If the individual component is missing, namely $\mu_i = 0$, pooled ols estimation is the most efficient estimator [@Croissant2008]. Under the assumptions of the FE model, the random effects estimators are biased and

[1] Five assumptions are required to produce consistent estimators with OLS : (i) a random sample of observations on y and $(x_1, ..., x_n)$, (ii) a random sample of n observations, (iii) no linear relationship among the explanatory variables, (iv) an error term that is uncorrelated with each explanatory variables and (v) an error term with zero mean conditional on the explanatory variables.

inconsistent, because μ_i is omitted and potentially correlated with the other regressors [@Schmidheiny2015]. Under the assumptions of the RE model, both the random and fixed effects estimation can be used.

Econometric Model

Consequently this study use equation 5 to test the combined effect of process and outcome-based CEP on CFP (short term vs long term).

$$\begin{aligned} Y_{it+1} = & \beta_0 + \beta_1(SPL_{it}) \\ & + \beta_2(STC_{it}) + \beta_3(A_{it}) \\ & + \beta_4(CaP_{it}) + \beta_5(WatP_{it}) \\ & + \beta_6(WastP_{it}) + (Controls_{it}) \\ & + \varepsilon_{it} \end{aligned} \tag{5}$$

where Y_{it+1} is a proxy of CFP measured as ROA (i.e. Model 1) or Tobin's Q (i.e. Model 2), SPL_{it} is a proxy for a firm's sustainability pay link, STC_{it} is a proxy for a firm's sustainability themed commitment, A_{it} is a proxy for a firm's audit score, EP_{it} is a proxy for a firm's energy productivity, CP_{it} is a proxy for a firm's carbon productivity, $WatP_{it}$ is a proxy for a firm's water productivity, $WasP_{it}$ is a proxy for a firm's waste productivity, $Controls_{it}$ is a vector of control variables that includes firm size, industry sector, financial leverage and growth and lastly ε_{it} which is the error term.

FE vs RE based on [@Bell2015] –> développez + appliquer sa méthode?

Panel data setting implies that endogeneity occurs in cases where the independent variable in a regression model is correlated with the error term, or due to simultaneous causality between the dependent and the independent variable [@Sanchez-Ballesta2007, @Biorn2008, @Roberts2013]. Consequently, the presence of endogeneity implies that the fourth and fifth assumptions of OLS[2] are violated and scholars have to use a different method to produce consistent estimators [@Wooldridge2008, @Roberts2013]. Recent meta-analysis provided evidences that the CFP-CEP nexus is characterized by a bidirectional causality [@Orlitzky2001, @Orlitzky2003, @Wu2006, @Albertini2013, @Dixon-Fowler2013, @EndrikatMakingsenseconflicting2014, @Ludecadedebatenexus2014, @WangMetaAnalyticReviewCorporate2016, @Busch2018]. In order to adress potential endogeneity problems in my model, firstly, I have lagged observations in dependent and control variables one year behind financial performance. This method allows to increase the confidence of the direction of the relationship [@Hart1996, @Delmas2015, @MiroshnychenkoGreenpracticesfinancial2017] and *in fine* reduce the potential simultaneity bias.

Secondly, given that the standard Hausman test had rejected the null hypothesis of random effect (see Annex... for results of the test or find a way to insert p-value in the table of regression idem for cross sectionnal dependence) I use a fixed effect model to regress the equation 5. According to @Roberts2013, fixed effect model improve endogeneity concerns.

Outliers treatment

@Lyu2015 defines outliers as observations in the dataset that appear to be unusual and discordant and which could lead to inconsistent results. @Osborne2004 have shown that even a small proportion of outliers can significantly affect simple analyses (i.e. t-tests, correlations and ANOVAs). Outliers are an issue only and only if they are influential [3] [@Cousineau2010]. I have used the Cook's distance [@Cook1977] test which is a common statistical tool to assess the influence of outliers [@JPStevens1984, @Cousineau2010, @Zuurprotocoldataexploration2010]. Cook's Distance observe the difference between the regression paramater

[2] Five assumptions are required to produce consistent estimators with OLS : (i) a random sample of observations on y and $(x_1, ..., x_n)$, (ii) a mean zero error term, (iii) no linear relationship among the explanatory variables, (iv) an error term that is uncorrelated with each explanatory variables and (v) an error term with zero mean conditional on the explanatory variables.

[3] Influential obervations are observations whose removal causes a different conclusion in the analysis

of a given model, $\hat{\beta}$ and what they become if the i_{th} data points is deleted,let's say $\hat{\beta}_i$. One difficulty with treatment of outliers is that the literature have not found common theoretical framework yet for the treatment of influential outliers [@OrrJohn1991, @Cousineau2010]. @Tabachnick2007 argue that the imputation with the mean is the best method while @Cousineau2010 highlights that it tends to reduce the spread of the population, make the observed distribution more leptokurtic, and possibly increase the likelihood of a type-I error. @Dang2009 argues that more elaborate technique involves replacing outliers with possible values while @Barnett1994 would prefere to remove or windsorized them. Alternatively, @Pollet2017 propose an other route to handle outliers and argue that inclusion or exclusion of outliers depend on the significativity of the results, meaning that if results are more significant without outliers, scholars should remove them and vice versa.

Following the mindset of @Pollet2017, I have concluded that model 1 using ROA as CFP proxies give better results with outliers and model 2 using Tobin's Q as CFP proxies give better results without outliers. See annex outliers for furthers details.

Sensitivity Analysis

Take ROE as another proxy of short term CFP. I need to find an other proxy for market-based indicator. I will also consider ESG factor of yahoo finance as a proxy for CEP.

To be continued...