Methodology

Panel Data : a theoretical background

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 offers 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 independent 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 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. 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. FE Model violates the fourth assumption while RE model implies that the common error component over individuals induces correlation across the composite error terms making the third assumption violated [@Croissant2008].

While OLS is not consistent to estimate panel data model, the R package *plm* provides pertinent estimation methods. (i) *The 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*, also called *within estimation*, transforms the original equation 1 in substracting the time averages to every variables, such as :

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

The presence of RE model in a panel data can be tested using the Breusch-Pagan Lagrange multiplier test [@Breusch1980]. This test is represented by the *plmtest* function in *R* and examines if time and/or individual specific variance components equal zero [@Park2011]. If Ho is verified, then there is no RE model in the panel data. The presence of FE model is tested by a F test (i.e. the function *pFtest* in *R*). The latter tests the individual and/or time effects based on the comparison of the within and the pooling model [@Croissant2008]. If Ho is verified then there is no FE model in the panel data.

In case of the absence of both RE and FE model, 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 inconsistent as μ_i is omitted and potentially correlated with the other regressors. Consequently, the fixed effects estimation need to be used. Under the assumption of the RE model, both FE and RE estimators are unbiased and consistent. According to @Schmidheiny2015, scholars should prefer the RE estimator only and only if $E[\mu_i, X_i] = 0$. This is tested by the Hausman test [@Hausman1981]. If Ho is verified then scholars should use RE estimator.

Econometric Model

This study uses the equation 5 to analyze the link between outcome-based and process-based CEP and the equation 6 to test the combined effect of process and outcome-based CEP on CFP (short term and long term).

$$Y_{it} = \alpha + \beta_1(SPL_{it}) + \beta_2(STC_{it}) + \beta_3(A_{it}) + u_{it} \tag{5}$$

where Y_{it} is a proxy of outcome-based CEP measured as carbon productivity (i.e. Model 1), water productivity (i.e. Model 2) and waste productivity (i.e. Model 3), 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 and u_{it} which is the error term.

$$\begin{aligned} Y_{it+1} = \quad & \alpha + \beta_1(SPL_{it}) \\ & +\beta_2(STC_{it}) + \beta_3(A_{it}) \\ & +\beta_4(CaP_{it}) + \beta_5(WatP_{it}) \\ & +\beta_6(WastP_{it}) + (Controls_{it}) \\ & +u_{it} \end{aligned} \tag{6}$$

where Y_{it+1} is a proxy of CFP measured as ROA (i.e. Model 4) or Tobin's Q (i.e. Model 5), 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, 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 u_{it} which is the error term.

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]. This could cause simultaneous causality between the dependent and the independent variable and lead to endogeneity concern in my model [@Sanchez-Ballesta2007, @Biorn2008, @Roberts2013]. In order to adress this issue I have lagged observations in dependent and control variables one year behind financial performance. This increases the confidence of the direction of the relationship [@Hart1996, @Delmas2015, @MiroshnychenkoGreenpracticesfinancial2017] and *in fine* reduce the potential simultaneity bias.

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 [1] [@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 observes the difference between the regression paramater 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 argues that the imputation with the mean is the best method while @Cousineau2010 highlight that it tends to reduce the spread of the population, making the observed distribution more leptokurtic, and possibly increase the likelihood of a type-I error. @Dang2009 argue that more elaborate technique involves replacing outliers with possible values while @Barnett1994 would prefer to remove or windsorized them. Alternatively, @Pollet2017 propose an

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

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 removed outliers from my database. See annex outliers for furthers details.