Methodology

Here is my methodology...

Panel Data

Definition of panel data

Panel data, also called longitudinal data or cross-sectional time-series data include observations on N cross section units (i.e., firms) over T time-periods.

Advantages of panel data :

As panel data analysis uses variation in both these dimensions, it is considered to be one of the most efficient analytical methods for data [@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].

Fixed or random effect model

Panel data may have individual (group) effect, time effect, or both, which are analyzed by fixed effect and/or random effect models. *A fixed effect model* examines if intercepts vary across group or time period, whereas a *random effect model* explores differences in error variance components across individual or time period. [@Park2011].

!! I need to test the fixed-random effect model of my database before moving forward !!

- @Ng2015 used the two-stage-least-square regressions to estimate its models.

** In case of presence of endogeneyity in an econometric model, OLS is not capable of delivering consistent parameter estimates [@Wooldridge2008].**

Citation from [@Wooldridge2008] :

> The general concept is that of the instrumental variables estimator; a popular form of that estimator, often employed in the context of endogeneity, is known as two-stage least squares (2SLS)

Endogeneity test

Even if panel data have a lot of advantages...

Two issues involved in utilizing panel data, namely heterogeneity bias and selectivity bias [@HsiaoChapitrePanelData2014].

Citation from @HsiaoChapitrePanelData2014:

> It is only by taking proper account of selectivity and heterogeneity biases in the panel data that one can have confidence in the results obtained.

@Dangsearchrobustmethods2015 examine which methods are appropriate for estimating dynamic panel data models in empirical corporate finance,especially in short panels of company data, in the likely presence of (1) unobserved heterogeneity and endogeneity, (2) residual serial correlation, or (3) fractional dependent

variables. The bias-corrected fixed-effects estimators, based on an analytical, bootstrap, or indirect inference approach, are found to be the most appropriate and robust methods.

But @MiroshnychenkoGreenpracticesfinancial2017 used the OLS regressions in micro panel using the Huber-White sand which estimator, to account for the heteroscedasticity problem... **Which method should I use?**

Hausmann test to test the random effects model for both dependent variables?

Econometric Model

The first hypothesis will be tested with T-tests on the impact of each green initiative on green performance.

Hypotheses two and three will be tested by regression analysis using the *plm* package. Econometric models are based on @Delmas2015 and @MiroshnychenkoGreenpracticesfinancial2017 and started from the general form:

$$Y_{t+1} = \beta_0 + \beta_1(X_{it}) + +\beta_2(C_{it}) + \varepsilon_{it} \tag{1}$$

where Y_{t+1} is the financial performance of firm i in year $t+1$, β is the vector of estimated regression coefficients for each of the explanatory variables X_{it}, C_{it} is a vector of control variables , ε_{it} is the error term.

More precisely I will regress six models :

Model 1 : Green Initiatives on Tobin's Q

$$TobinsQ_{it+1} = \beta_0 + \beta_1(SP_{it}) + \beta_2(ST_{it}) + \beta_3(AS_{it}) + \beta_4(C_{it}) + \varepsilon_{it} \tag{2}$$

Model 2 : Green Initiatives on ROA

$$ROA_{it+1} = \beta_0 + \beta_1(SP_{it}) + \beta_2(ST_{it}) + \beta_3(AS_{it}) + \beta_4(C_{it}) + \varepsilon_{it} \tag{3}$$

Model 3 : Green Performance on Tobin's Q

$$TobinsQ_{it+1} = \beta_0 + \beta_1(EP_{it}) + \beta_2(CP_{it}) + \beta_3(WatP_{it}) + \beta_4(WasP_{it}) + \beta_5(C_{it}) + \varepsilon_{it} \tag{4}$$

Model 4 : Green Performance on ROA

$$ROA_{it+1} = \beta_0 + \beta_1(EP_{it}) + \beta_2(CP_{it}) + \beta_3(WatP_{it}) + \beta_4(WasP_{it}) + \beta_5(C_{it}) + \varepsilon_{it} \tag{5}$$

Model 5 : Both Green Performance and Green Initiative on Tobin's Q

$$TobinsQ_{it+1} = \beta_0+\beta_1(EP_{it})+\beta_2(CP_{it})+\beta_3(WatP_{it})+\beta_4(WasP_{it})+\beta_5(SP_{it})+\beta_6(ST_{it})+\beta_7(AS_{it})+(C_{it})+\varepsilon_{it} \tag{6}$$

Model 6 : Both Green Performance and Green Initiative on ROA

$$ROA_{it+1} = \beta_0+\beta_1(EP_{it})+\beta_2(CP_{it})+\beta_3(WatP_{it})+\beta_4(WasP_{it})+\beta_5(SP_{it})+\beta_6(ST_{it})+\beta_7(AS_{it})+(C_{it})+\varepsilon_{it} \tag{7}$$

where :

- $TobinsQ_{it+1}$ = a proxy for a firm's financial performance
- ROA_{it+1} = a proxy for a firm's financial performance
- EP_{it} = a proxy for a firm's energy productivity
- CP_{it} = a proxy for a firm's carbon productivity
- $WatP_{it}$ = a proxy for a firm's water productivity
- $WasP_{it}$ = a proxy for a firm's waste productivity
- SP_{it} = a proxy for a firm's sustainability pay link
- ST_{it} = a proxy for a firm's sustainability themed commitment
- EP_{it} = a proxy for a firm's audit score

- C_{it} = a vector of control variables that include financial leverage, firm size, net margin and industry sector
- ε_{it} = the error term

Panel Data Tests

This section will not be in the final document. It is only to report the result of the bunch of tests I carried out in order to define which panel data methotolodies I will use for each one of my 6 models.

@Croissant2008a and @Torres-Reyna2010 really helped me.

Here are the tests :

1. Test of poolability
2. Hausmann Test to determine the fixed or random effect
3. Test for time fixed effect
4. Test for cross-sectional dependence
5. Test for serial correlation
6. Test for stationarity
7. Test for heteroskedasticity

The table 1 summaries the result of each test for each model. You can find details below.

Regarding the poolability test I have an issue with my code that I still need to solve. This is why it is written *NA* in the table 1. I have also an issue with the test for cross-sectionnal dependence. Indeed depending the method I use with the test syntax (i.e. Pesaran's CD test (test="cd") or Breusch and Pagan's LM test (test="lm"), I got divergent results. **Do you know why? Which one suit the best my model?**

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Poolability	NA	NA	NA	NA	NA	NA
Hausmann	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed
Time Fixed Effect	No	Yes	No	Yes	No	Yes
Cross Sectional Dependence	?	Yes	?	?	?	Yes
Serial Correlation	Yes	Yes	Yes	Yes	Yes	Yes
Stationarity	None	None	None	None	None	None
Heteroskedasticity	Yes	Yes	Yes	Yes	Yes	Yes

Table 1: Test Summary

Test of poolability

Citation from [@Croissant2008] :

> *Pooltest tests the hypothesis that the same coefficients apply to each individual. It is a standard F test, based on the comparison of a model obtained for the full sample and a model based on the estimation of an equation for each individual. The first argument of pooltest is a plm object. The second argument is a pvcm object obtained with model=within. If the first argument is a pooling model, the test applies to all the coefficients (including the intercepts), if it is a within model, different intercepts are assumed.*

The null hypothesis of poolability assumes homogeneous slope coefficients.

When running my code I got this error : Error in FUN(X[[i]], ...) : insufficient number of observations

I still need to understand the origin of this error.

Hausmann Test to determine the fixed or random effect

Citation from @Torres-Reyna2010 :

> *To decide between fixed or random effects you can run a Hausman test where the null hypothesis is that the preferred model is random effects vs. the alternative the fixed effects (see Green, 2008, chapter 9). It basically tests whether the unique errors (ui) are correlated with the regressors, the null hypothesis is they are not.*

The Table 2 summarizes results of the Hausman Test of each model. We can observe that all p-values are $<$ 0.05 meaning that HO is not verified and all models are caraterized by a fixed effect.

Table 2: Hausman Test PValue

Model	P-Value
Model 1	3.91743371664877e-11
Model 2	0.00295804024618629
Model 3	2.03716389543958e-08
Model 4	6.48087009803431e-06
Model 5	4.61015773467216e-08
Model 6	7.7661907780088e-07

Test for time fixed effect

The Table 3 summarizes results of the test for each model.

P-Value is > 0.05 for model 1, model 3 and model 5 meaning that null hypothesis is verified and that there is not a significant time-fixed effect. However for model 2,model 4 and model 6 P-Value is < 0.05 meaning that null hypothesis is rejected and that there is a significant time-fixed effect.

Does this mean that for model 2,4 and 6 I have to add the time fixed effect in my model?

Table 3: Fixed Time Effect Test PValue

	Model	P-Value
F	Model 1	0.294413678243895
F	Model 2	0.000368808643889729
F	Model 3	0.430605654981738
F	Model 4	0.0012952612768481
F	Model 5	0.563399152332159
F	Model 6	0.000818153246924005

Test for cross-sectional dependence

Citation from @Torres-Reyna2010 :

> "According to Baltagi, cross-sectional dependence is a problem in macro panels with long time series. This is not much of a problem in micro panels (few years and large number of cases). The null hypothesis in the B -P/LM and Pasaran CD tests of independence is that residuals across entities are not correlated. B- P/LM and Pasaran CD (cross-sectional dependence) tests are used to test whether the residuals are correlated across entities*. Cross-sectional dependence can lead to bias in tests results (also called contemporaneous correlation)."

Table 4: Test for cross-sectional dependence - Result PValue

Model	Method	P-Value
Model 1	cd	5.1349772167113e-06
Model 2	cd	6.47407516580982e-08
Model 3	cd	0.00395522799662467
Model 4	cd	0.139271504054302
Model 5	cd	0.312716750944021
Model 6	cd	0.0603296334211678

The method 'cd' stands for Pesaran's CD Statistic

Sensitivity Analysis