Are international environmental agreements stable ex-post?

Beard, Rodney and Mallawaarachchi, Thilak

Group Sup de Co La Rochelle France, The University of Queensland, School of Economics

25 October 2011

Are international environmental agreements stable ex-post?[1]

Rodney Beard
Groupe Sup de Co La Rochelle
102 rue de coureilles
Les Minimes
La Rochelle 17024 Cedex 1
France

And

Thilak Mallawaarachchi
Risk and sustainable management group
School of Economics
The University of Queensland
Brisbane 4072
Australia

[1] An earlier version of this paper was presented at the "International Forum - Towards Global Agreements on Environmental Protection and Sustainability: Frontiers of Future Economic Research" University of Exeter, Wednesday 13th - Friday 15th April 2011, and also at the European association for Environmental Resource Economics Meeting in Rome July 2011. We would like to thanks participants at those events for discussions and comments.

Abstract

In this paper we present a model of international environmental agreements in the presence of threshold effects. The model is in the tradition of models of international environmental agreements formulated as games in partition function form. Games in partition function form allow the incorporation of external effects between players. The model is applied to global climate change agreements. The agreement involves a contract between nations as to the level of abatement of greenhouse gas emissions and how these benefits are to be shared. Benefits to emissions abatement are subject to a threshold. Consequently, we model climate as a global threshold public good. This allows a mechanism to explore incentives and disincentives for signing agreements consequent to a critical number of other players committing to an agreement. We show that thresholds may destabilize what would be an otherwise stable agreement and that combining an emissions tax with an international agreement can be used to restore stability.

1. Introduction

Climate change, water security, and the loss of biological diversity are some of the most important environmental problems that undermine the sustainability of the modern interconnected economies. All these problems share a common thread in that uncoordinated individual actions are contributing to the depletion of commonly held property that form an integral part of the global natural environment. The uncertainties relating to the nature of cause and effect and the inability to hold individuals to account for damage makes it necessary to reach a solution through cooperation. International Environmental Agreements (IEAs) negotiated between nations are an example of a global policy instrument designed to improve global welfare. The challenge in the IEAs has been to make them not only profitable, but also self-enforcing, due in part for incentives for nation to join and remain committed in their own self-interest (Fuentes-Albero, & Rubio 2010).

The literature on the minimum number of players needed to form an effective international environmental agreement (IEA) includes Cararro, et al. (2004). In their work, the participation problem is formulated as a three-stage game where players choose the minimum proportion of the total number of players who must be signatories in the first stage of the game. In this paper, we argue that it is not so much the number of players but their contribution to emissions or abatement is the central issue. We therefore model IEA as a game in partition function form with a threshold or provision point public bad, namely greenhouse gas emissions. Games in partition function form were first proposed by Thrall and Lucas (1963) and later applied to strategic cooperation in international environmental agreements), for example by Chander and Tulkens (1997) and, McQuillin (2008)

Barratt (1994) also uses a similar set-up although in his model Chander and Tulkens' damage function takes the form of a benefit function, and emissions are replaced with abatement levels. In that model, the transformation function becomes a cost function. The differences, however, do not appear to be essential in informing the outcome.

In this paper, we examine how a threshold or provision point affects the partial agreement equilibrium. In other words, does a high threshold - an emissions level that is unlikely to be reached, leading to a smaller group of signatories in equilibrium? Then, can a partial agreement equilibrium induce emissions levels that drive the ambient greenhouse gas concentration below the provision point? This is really the key practical question, because if this is not the case, then IEA will be ineffective in eradicating damage. Perhaps eradication of damage is wishful thinking and in a world where damage is continuous and thresholds don't play a role this is indeed the case. On the other hand, we may not have exceeded a critical threshold, or if we have, we may not have exceeded by too much, and still through sensible management practices, able to push ambient greenhouse gas concentrations back below some threshold. An important question is: Are IEA with a limited number of signatories able to do this? We examine this question below.

2. The Model

The model is based on Chander (2007) and Chander and Tulkens (1997). We consider a game between N players or countries where N={1,…,n}. Furthermore, we will consider a partition of this set such that: $P=\left(S_1,\ldots,S_m\right)$ and $\bigcup_{j=1}^{m} S_j = N$. Where, for all

$i \neq j, S_i \cap S_j = \varnothing$ and we will refer to this partition as a coalition structure.

Commodities are of two types: a private good $y_i, i = 1, \ldots, n$ and a public bad z which represents the ambient level of greenhouse gas concentration in the atmosphere (e.g. CO_2)[2]. The private good y_i is related to greenhouse gas emissions e_i according to the following transformation function:

$$y_i = g(e_i), i \in N \text{ and } z = \sum_{i \in N} e_i \tag{1}$$

Our model then departs somewhat from the Chander and Tulkens (1997) framework by introducing the idea that environmental damage (i.e. climate change) should be considered a threshold public good. In other words emissions only induce damage when the ambient level of greenhouse gases exceeds some threshold which in the threshold public goods literature is typically referred to as a provision point. Consequently we model consumer preferences in terms of a provision point mechanism:

$$u_i(y_i, z) = y_i - v(z), \text{ if } z > z_0 \text{ and } u_i(y_i, z) = y_i \text{ if } z \leq z_0 \tag{2}$$

Where the provision point z_0 may be interpreted as a threshold below which total emissions are insufficient to induce environmental damage in the sense of global warming.

We now proceed as follows: first, we determine the Pareto efficient allocation in each of these cases before defining the γ-characteristic function and γ-core of the game. Then we analyze for each case the conditions under which a partial agreement equilibrium (a particular type of Nash equilibrium for partition function form games) exists before proceeding to study how the existence of a "provision point" may impact on the Chander-Tulkens solution to the game.

[2] Because the model is essentially static ambient C02 concentration just equals the sum of the emissions.

If $\left(e_1^*,\ldots e_n^*\right)$ are the Pareto efficient emissions then for the first case the first order conditions are given by $g'\left(e_i^*\right)=nv'(z)$and in the second case one obtains $g'\left(e_i^*\right)=0$. From the latter condition we can deduce that $\left.\frac{dv}{dz}\right|_{z=z_0}=0$ for some $n>0$. These conditions are interesting because, assumption 4 of Chander and Tulkens (1997) allows for threshold public goods but they do not analyze the formation of international environmental agreements in terms of this threshold but only in order to set-up the damage function. In their later work they drop this specification and assume away a provision point - this allows them to weaken the assumptions they require regarding concavity of the transformation function.

We now consider the case where emissions are above the threshold in more detail. In terms of a strategic game first we define the strategy space $T_i=\left\{e_i : 0\le e_i\le e^0\right\}$, and $T=T_1\times\ldots\times T_n$ and utility profile $u=\left(u_1,\ldots u_n\right)$. This defines a game $\Gamma=\left(N,T,u\right)$. Denote the Nash equilibrium of this game by $\left(\bar{e}_1,\ldots\bar{e}_n\right)$. Given a coalition structure P we can define a coalitional equilibrium as

$$\left(\bar{e}_i\right)_{i\in S_j}=\arg\max\left(\sum_{i\in s_j}\left[g\left(e_i\right)-v\left(\sum_{i\in S_j}e_i+\sum_{k\in N\setminus S_j}\bar{e}_k\right)\right]\right), j=1,2,\ldots,m$$

Note that this implies that $\sum_{i\in S_j}e_i+\sum_{k\in N\setminus S_j}\bar{e}_k>z_0$.

We now consider second case in which emissions are below the threshold in more detail. In this case the coalitional equilibrium plays no role and one obtains $g'\left(\bar{e}_i\right)=0, i=1,...,n$, in other words the game theoretic character of the problem disappears and each country unilaterally determines emissions level as they see fit. Their emissions have no impact on each-other.

3. The γ-characteristic function

We now consider a partition consisting of a coalition S and a number of individual players, the coalitional equilibrium now takes the form (partial agreement equilibrium):

$$(\hat{e}_i)_{i\in S} = \arg\max\left(\sum_{i\in s}\left[g(e_i)-v\left(\sum_{i\in S}e_i+\sum_{j\in N\setminus S}\hat{e}_j\right)\right]\right)$$

And

$$\hat{e}_j = \arg\max\left[g(e_j)-v\left(\sum_{i\in N, i\neq j}\hat{e}_i+e_j\right)\right], j\in N\setminus S.$$

The first of these gives the conditions under which the coalition S maximizes welfare and the second gives the best-response of a non-member of the coalition to the optimal emissions decisions of the coalition. We now introduce the γ-characteristic function:

$$w^{\gamma}(S)\equiv\sum_{i\in S}\left[g(\hat{e}_i)-v\left(\sum_{j\in N}\hat{e}_j\right)\right].$$

Therefore the γ-characteristic function gives an expression for the surplus welfare generated by coalition members. In the event that emissions are constrained to not exceed the threshold the disutility term in each of these expressions will be zero. This completes the set-up of the model.

The rest of our paper considers the implications of thresholds for proposition 5 of Chander and Tulkens (1997). Proposition 5 is here re-stated without proof in slightly modified form.

Proposition 1: *For all $S\subset N, S\neq N, |S|\geq 2, n_s v'(z^*)\geq v'(\bar{z})$, where n_S denotes the size of coalition S and corresponds to the Nash (disagreement) equilibrium $\bar{z}$ and Pareto efficient levels of ambient emissions z^* respectively. Then the emission level of each player in the coalition of a partial agreement equilibrium is not higher than the emission level corresponding to the Nash equilibrium.*

The result as stated here relies on symmetry of the disutility of ambient emissions. This assumption was not made in Chander and Tulkens (1997) but is made in later work, e.g. Chander (2007).

Our first result is a corollary of this. This result is not really surprising and rather obvious however it is presented because it will be referred to later.

Proposition 2: *Given the validity of proposition 1 there is some minimal size of the coalition that guarantees at least one signatory to the agreement.*

Proof: Proposition 1 implies that $z^* \leq \hat{z} \leq \bar{z}$, however, strict concavity of the damage function implies $v'(z^*) > v'(\bar{z})$ (see figure 1). This implies $n_S \geq \varepsilon < 1$. If z^* is less than the threshold then coalition size will be large as long as the disagreement equilibrium does not induce emissions that are too large. In other words if one is already too far above the threshold the coalition size will be small. ■

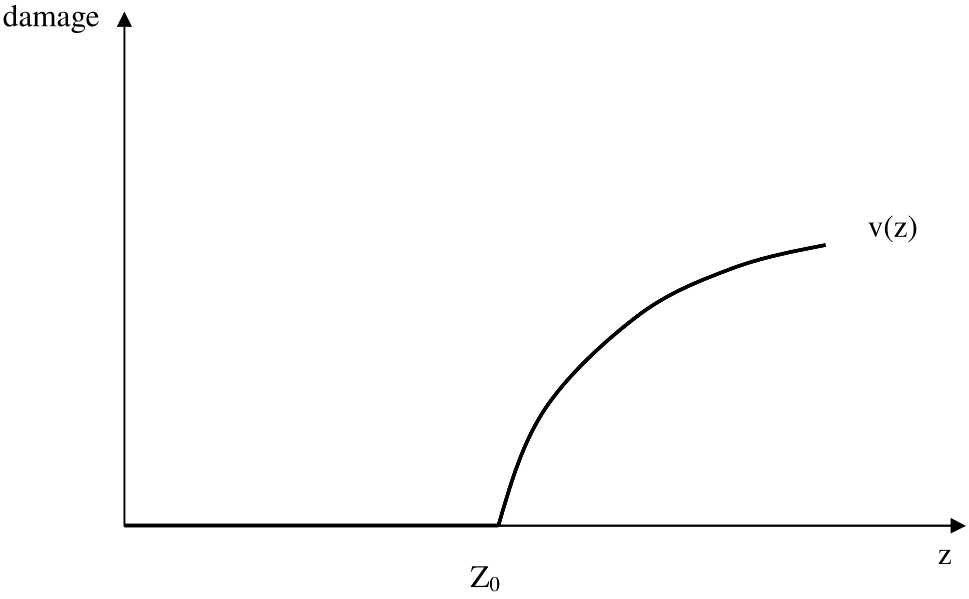

Figure 1: damage or disutility function

Because the harmful emissions threshold for 'dangerous global warming' is not precisely known, it is important to analyse how to reduce emissions if the coalition membership were to remain below the critical membership threshold?

Consider the following situation. Initially $\bar{z} > z_0$, however, as a result of the agreement emissions $\hat{z}$ are reduced to a level: $\hat{z} < z_0$. We assume here that emissions under the Pareto-optimal outcome are even lower. In other words, we assume both the potential for the success in terms of emission reductions of a full agreement and the success on the same terms of a 'partial agreement equilibrium'. This however has as a consequence that ex-post each player maximizes their own private benefits to emissions and that

$$g'(e_i^*) = g'(\hat{e}_i) = g'(\bar{e}_i) = 0, \forall i .$$

It is possible that this behaviour could induce an aggregate level of emissions that would be once again lie above the threshold. Therefore, any viable agreement requires an additional condition for it to be workable and effective in keeping emissions below the harmful threshold. We term this condition ex-post robustness. It is likely related to the idea of ex-post implementation and robust implementation in the mechanism design literature, however that literature is largely non-cooperative in nature rather than cooperative [see for example, Bergemann & Morris (2008, 2009)]. Firstly we define the ex-post emissions equilibrium as follows.

Definition 1: *Ex-post emissions equilibrium* **.**

$$(e_i^{ep})_{i \in N} = \arg\max g(e_i), \forall i \in N$$

Note that this is not the same as the disagreement equilibrium, which is defined in terms of damage from emissions. The ex-post equilibrium assumes that damaging levels of emissions have been eradicated although not necessarily that all emissions have been eradicated.

Definition 2: *Ex-post robustness. A partial agreement equilibrium* $(\hat{e}_i, \hat{e}_j)_{i\in S, j\in N\backslash S}$ *is said to be ex-post robust iff* $z^{ep} \leq z_0$.

If the agreement were not ex-post robust, then self-interested behavior of all parties whether signatories or not after successful reduction of emissions would lead to emissions again increasing to a new level above the provision point threshold which would trigger the need for a new agreement. Essentially the initial IEA would not be time consistent, although fully exploring time consistency properties would require development of a multi-period or fully dynamic model of IEA's (see recent work by Pavlova, 2008).

Example 1:

Consider the following utility function $u_i = ae_i - be_i^2 - c\sum_i e_i$. This utility function assumes a linear damage function above the threshold. This induces a partial agreement equilibrium as follows:

$$\hat{e}_{i\in S} = \frac{a - n_S c}{2b}, \hat{e}_{j\in N\backslash S} = \frac{a-c}{2b}.$$

The ex-post equilibrium is

$$e_i^{ep} = \frac{a}{2b}, i \in N.$$

From this it is clear that $\hat{z} = n_S \frac{a - n_S c}{2b} + (n - n_S)\frac{a-c}{2b} = n\frac{a}{2b} - n\frac{c}{2b} - \frac{(n_S^2 - 1)c}{2b}$. If the treaty is successful in reducing emissions this will be less than z_0. However, the ex-post equilibrium induced emissions are $z^{ep} = n\frac{a}{2b}$. Clearly, $z^{ep} > \hat{z}$. Nevertheless, there clearly could exist a z_0, such that $z^{ep} > z^0 > \hat{z}.$ So that the treaty, on being implemented, would create incentives for individuals to again pollute to damaging levels.

The threshold could however be even larger than the ex-post emissions level, which would make the success of the treaty even greater in terms of emissions reductions and return us to a state of the world predating the era of human induced climate change. More likely however is that an agreement will be moderately successful in controlling emissions because the threshold itself for inducing damage is not that great or we have not already passed it by much. A low threshold is easy to cross again, and this makes for IEA that are unlikely to be ex-post robust.

We now turn to the key question that we proposed in the introduction how in equilibrium is the number of signatories affected by the threshold level? Assuming a utility function $u_i = ae_i - be_i^2 - c\sum_i e_i$ like that of the previous example, then we can state the following proposition.

Proposition 3: $n_S \geq \sqrt{\frac{n(a-c) - 2bz_0}{c} + 1}$

Proof: a partial agreement equilibrium for the quadratic utility function implies $\hat{z} = n\frac{a-c}{2b} - \frac{(n_S^2 - 1)c}{2b} \leq z_0$ rearranging one obtains $n_S \geq \sqrt{\frac{n(a-c) - 2bz_0}{c} + 1}$ ■

Clearly, our initial intuition is valid, and in this case, a higher threshold for pollutants would in equilibrium result in a lower minimum threshold for membership of the coalition in an agreement that reduces emissions below damaging levels. This should be interpreted to mean that it would be hard to gain signatories to agreements if the damage threshold is high. Low thresholds which are easily crossed on the other hand and for which the success of emissions reductions is easily achieved correspond to situations for which it is easier to obtain commitment for a binding agreement. The downside of such agreements is that they may well not be robust as the previous example illustrates.

Next we explore how the conditions for internal and external stability of an agreement and the use of transfer payments in achieving stability.

4. Ex-post solution of a partial IEA

What consequences might thresholds have for transfers between members of a coalition? Consider the Chander and Tulkens transfer mechanism. This determines transfers between players that guarantee internal and external stability of the agreement. However, these transfers do not relate to the robustnesss of the agreement ex-post. Consequently, the Chander Tulkens solutions concept is an ex-ante solutions concept.

Chander-Tulkens transfers are defined as follows:

$$T_i = -\left(g_i\left(e_i^*\right) - g_i\left(\bar{e}_i\right)\right) + \frac{v'\left(z^*\right)}{nv'\left(z^*\right)}\left[\sum_{j\in N} g\left(e_j^*\right) - \sum_{j\in N} g\left(\bar{e}_j\right)\right]$$
$$= -\left(g_i\left(e_i^*\right) - g_i\left(\bar{e}_i\right)\right) + \frac{1}{n}\left[\sum_{j\in N} g\left(e_j^*\right) - \sum_{j\in N} g\left(\bar{e}_j\right)\right]$$

The transfer T_i is positive for a transfer received and negative for transfers made. These transfers act to compensate players for losses due to environmental damage. The C-T transfer mechanism essentially applies a proportional rule to distribute the surplus of the grand coalition along with the opportunity cost associated with unanimous agreement amongst all players. It is worth noting that because the good in question is a global public good, these transfers are paid to all players whether or not they are signing members of a coalition.

Note because we follow Chander (2007) in assuming that damage from climate change is a pure public bad and therefore identical for all players. The marginal damage does not appear in the Chander-Tulkens transfer formula. Consequently the threshold will not impact on transfers directly, but it will have an impact in equilibrium as the following example illustrates.

Example 2: Chander-Tulkens mechanism with threshold

Using the quadratic utility function from of the last example, however, with different parameters for different players, we can compute the Pareto efficient emissions, to be:

$$e_i^* = \frac{a_i - nc}{2b_i}, i \in N .$$

And the disagreement equilibrium is:

$$\bar{e}_i = \frac{a_i - n_S c}{2b_i}, i \in S .$$

This results in the following transfers if emissions are above the threshold.

$$\begin{aligned} T_i = {} & -\left(a_i \frac{a_i - nc}{2b_i} - b_i \left(\frac{a_i - nc}{2b_i} \right)^2 - \left(a_i \frac{a_i - n_S c}{2b_i} - b_i \left(\frac{a_i - n_S c}{2b_i} \right)^2 \right) \right) \\ & + \frac{1}{n} \left[\sum_{j \in N} a_i \frac{a_i - nc}{2b_i} - b_i \left(\frac{a_i - nc}{2b_i} \right)^2 - \sum_{j \in N} a_i \frac{a_i - n_S c}{2b_i} - b_i \left(\frac{a_i - n_S c}{2b_i} \right)^2 \right] \end{aligned}$$

Simplifying,

$$\begin{aligned} T_i = {} & -\left(a_i \frac{(n_s - n)c}{2b_i} + b_i \left(\left(\frac{a_i - n_S c}{2b_i} \right)^2 - \left(\frac{a_i - nc}{2b_i} \right)^2 \right) \right) \\ & + \frac{1}{n} \left[\sum_{j \in N} a_i \frac{(n_s - n)c}{2b_i} + b_i \left(\left(\frac{a_i - n_S c}{2b_i} \right)^2 - \left(\frac{a_i - nc}{2b_i} \right)^2 \right) \right] \end{aligned}$$

Can we place a bound on transfers based on the bound on coalition size? To do this we need to distinguish ex-post and ex-ante cases.

However c is zero if aggregate emissions are below the threshold. So we need to distinguish the following cases:

i. emissions below the threshold $T_i = 0$

ii. emissions above the threshold

$$T_i \geq -\left(a_i \frac{\left(\sqrt{\frac{n(a-c)-2bz_0}{c}}+1-n\right)c}{2b_i} + b_i \left(\left(\frac{a_i - \left(\sqrt{\frac{n(a-c)-2bz_0}{c}}+1\right)c}{2b_i} \right)^2 - \left(\frac{a_i - nc}{2b_i}\right)^2 \right) \right)$$
$$+\frac{1}{n}\left[\sum_{j\in N} a_i \frac{\left(\sqrt{\frac{n(a-c)-2bz_0}{c}}+1-n\right)c}{2b_i} + b_i \left(\left(\frac{a_i - \left(\sqrt{\frac{n(a-c)-2bz_0}{c}}+1\right)c}{2b_i} \right)^2 - \left(\frac{a_i - nc}{2b_i}\right)^2 \right) \right]$$

Therefore we only need to consider the latter case in which emissions are above the threshold.

In this case transfers do depend on the threshold. If we consider the lower bound of transfers, i.e. the minimum level of transfers necessary to maintain an agreement then we can examine how an increase in the threshold impacts transfers. This is of interest because it is unclear where such a threshold may lie. Estimates of what level of emissions are likely to be damaging vary.

As discussed earlier there clearly could exist a z_0, such that $z^{ep} > z^0 > \hat{z}$. So that the treaty, on being implemented, would create incentives for individuals to again pollute to damaging levels. The question we now pose is whether a policy can be devised to exclude this possibility? In other words can we design a policy that guarantees ex-post robustness of any international environmental agreement? Firstly, note that we will assume an ex-ante successful treaty in other word $\acute{z} < z_0$. However to guarantee ex-post robustness as we define it here we need ex-post emissions also to remain below the threshold. For the quadratic utility example ex-post emissions are given by $z^{ep} = n\frac{a}{2b}$. While n is clearly independent of policy. a, and b represent the marginal private benefit of emissions and $2b$ the marginal private cost of emissions. Lowering the marginal private benefit of emissions or raising the marginal private cost of emissions would lead to a reduction of ex-post emissions. If these can be reduced to a level below the damage threshold then we would have achieved our goal of reducing emissions below the threshold ex-post as well. The most obvious way of achieving this would be to couple negotiations on international environmental agreements to emissions taxes. This case will be considered in section 6.

A second issue raised by Fuentes-Albero and Rubio (2010) is the case of transfers between rich and poor countries that may possess different cost structures. For example in our set-up the cost of emissions damages may vary across countries if there were for example an income effect on peoples willingness to pay for abatement or alternatively willingness to accept various emissions levels. This leads to consider transfer payments with heterogenous damage costs. We do not pursue this here. Finally, we clarify the impact of threshold effects on the γ-core of the game.

5. Thresholds and the γ −core

In this section we analyse the γ −core with thresholds and examine the impact of thresholds on the γ −core. In order to characterize the γ −core we first need to introduce some definitions which we have hitherto not required.

Definition 3 *Feasible states are vectors*

$$(y, e, z) \equiv (y_1, .., y_n; e_1, .., e_n; z)$$

Such that

$$\sum_{i \in N} y_i \leq \sum_{i \in N} g_i(e_i)$$

And

$$z = \sum_{i \in N} e_i$$

Definition 4 *A Pareto efficient state of the economy is a feasible state (y, e, z) such that there exists no other feasible state (y', e', z') for which $u_i(y_i', z') \geq u_i(y_i z)$ for all $i \in N$ with strict inequality holding for at least one i.*

The procedure employed by Chander and Tulkens to demonstrate non-emptiness of the $\boldsymbol{\gamma}$ −core is to first define a strategy for the game and then show that this lies in core of game. An alternative procedure was employed by Helm who extended the Bondareva-Shapley theorem to a game with multilateral externalities.

Definition 5 *A strategy of the coalition N is said to belong to the core of the corresponding game (N, w) if the the payoff it yields for each coalition is larger than the payoff $w(S)$ that any coalition $S \subset N$ can achieve.*

Chander and Tulkens proved non-emptiness of the core for the case of symmetric production functions and asymmetric non-linear damage functions. Here we consider the case of asymmetric production functions and symmetric damage functions. We follow Helm (2001) with some modifications based on Osborne and Rubinstein's (1994) presentation of the Bondareva-Shapley theorem.

Definition 6 *A game is said to be balanced iff $\sum_{S \subset C} \delta_S w(S) \leq v(N)$ for all balanced collection of weights, where $(\delta_S)_{S \in C}$ is a balanced collection of weights if the sum of δ_S summed over all coalitions S containing I is 1.*

Proposition 4 *For $z \leq z_0$ the core of the game is ex-post empty.*

Proof: To prove this we begin by using Helm's approach to proving non-emptiness of the $\boldsymbol{\gamma}$ −core. This applies the Bondareva-Shapley theorem to games with multilateral environmental externalities. We begin by assuming that the game is balanced and therefore

$v(N) \geq \sum_S \delta_S w^\gamma(S)$

First recall that the the γ-characteristic function is given by

$$w^\gamma(S) \equiv \sum_{i \in S}\left[g(\hat{e}_i) - v\left(\sum_{j \in N} \hat{e}_j\right)\right].$$

And that in the event of emissions falling below the threshold the damage function will drop-out. In this case the characteristic function becomes

$$w^\gamma(S) = \sum_{i \in S} g(\widehat{e_i})$$

Ex_post robustness means that $z^{ep} \leq z_0$ and that damages remain at zero. If the game is balanced ex-ante and the core is non-empty then ex-ante a successful treaty results in

$$\sum_{S \in C} \delta_S w^\gamma(S) \leq \sum_{S \in C} \delta_S \sum_{i \in N} g(\widehat{e_i}) = \sum_{i \in N} g(\widehat{e_i}) \sum_{S \ni i} \delta_S = \sum_{i \in N} g(\widehat{e_i}) = v(N)$$

In the ex-post non-robust case the characteristic function becomes

$$w^\gamma(S) = \sum_{i \in S}\left[g\left(e_i^{ep}\right) - v\left(\sum_{j \in N \in} e_j^{ep}\right)\right]$$

If the ex-post equilibrium induces an outcome in the core then this results in the following balanced game condition

$$\sum_{S\in C}\delta_S w^{\gamma}(S)$$

$$\leq \sum_{S\in C}\delta_S\sum_{i\in S}\left[g\left(e_i^{ep}\right)-v\left(\sum_{j\in N\in}e_j^{ep}\right)\right]$$

$$=\sum_{i\in N}\left[g\left(e_i^{ep}\right)-v\left(\sum_{j\in N\in}e_j^{ep}\right)\right]\sum_{S\ni i}\delta_S=\sum_{i\in N}\left[g\left(e_i^{ep}\right)-v\left(\sum_{j\in N\in}e_j^{ep}\right)\right]>\sum_{i\in N}g(\widehat{e_\imath})$$

The last step is true because otherwise there would be no-incentive to move from the ex-ante position to the ex-post position. In other words you would not deviate from an agreement ex-post unless the net gain from deviating were greater than that from the payoff you would receive from adhering to the agreement. This will be the case if the slope of the production function is greater than that of the damage function and this will occur only if emissions lie below the Pareto optimal level but above the threshold. Which is by proposition 1 impossible so that we conclude the postulated outcome is not in the core. To see this recall the first-order conditions for a Pareto optimim in this model $g'(e)=nv'(z)$ and proposition 1.

The result is that after dropping intermediate terms:

$$\sum_{S\in C}\delta_S w^{\gamma}(S)>v(N)$$

In the ex-post robust case the balanced game condition gives

$$\sum_{S\in C}\delta_S w^{\gamma}(S)\leq\sum_{S\in C}\delta_S\sum_{i\in N}g\left(e_i^{ep}\right)=\sum_{i\in N}g\left(e_i^{ep}\right)\sum_{S\ni i}\delta_S=\sum_{i\in N}g\left(e_i^{ep}\right)>\sum_{i\in N}g(\widehat{e_\imath})=v(N)$$

Consequently the game is not balanced.•

Note that non-coalition members would also increase emissions under these circumstances as they are not penalized by environmental damage if emissions fall below the threshold. However in the ex-post equilibrium there is really no asymmetry in behavior between signatories and non-signatories of an agreement because each acts in their own self interest and strategic considerations do not play a role (the game theoretic aspect of the problem disappears when there are no damages).

Threshold effects raise questions about the robustness of IEA's that are a concern if post-implementation of the agreement emissions rise again to damaging levels, this may or may not be a problem depending on the specific incentives of individual countries. While stable agreements seem possible in world without thresholds the possible existence of thresholds can create incentives to deviate from successful agreements.

6. Can an emissions tax stabilize an agreement ex-post?

An emissions tax such as a carbon tax would raise the marginal private cost of emissions. A carbon tax might therefore complement transfer and compensation policies. How high would such a tax need to be in order to lead to lasting stable agreements? To answer this question, we reformulate the above model by incorporating an emissions tax. We develop the analysis by way of an example. Consider the utility function with linear damages above the threshold:

$$u_i = ae_i - be_i^2 - te_i - c\sum_{i=1}^{n} e_i$$

Where t is now an emissions tax such as a carbon tax per unit of emissions. Now consider the partial agreement equilibrium in the presence of a carbon tax on emissions:

$$(\hat{e}_i)_{i\in S} = \arg\max\left(\sum_{i\in s}\left[g(e_i) - te_i - v\left(\sum_{i\in S} e_i + \sum_{j\in N\setminus S}\hat{e}_j\right)\right]\right)$$

And

$$\hat{e}_j = \arg\max\left[g(e_j) - v\left(\sum_{i\in N, i\neq j}\hat{e}_i + e_j\right)\right], j\in N\setminus S\,.$$

The assumption here is that the agreement now consists of an emissions agreement and a harmonized emissions tax. Consequently non-signatories do not incu a tax. Admittedly one could conceive of a situation where a country prefers a tax to an emissions agreement and therefore does not sign the emissions agreement but does impose a tax. We do not consider this situation.

For the utility function that is considered here by way of example the partial agreement equilibrium would be as follows:

$$\left(\hat{e}_i\right)_{i\in S} = \frac{a-t-n_S c}{2b}$$

and

$$\hat{e}_{j\in N\backslash S} = \frac{a-c}{2b}$$

The ex-post equilibrium with tax is given by

$$\left(e_i^{ep}\right)_{i\in S} = \frac{a-t}{2b}$$

and

$$e_{j\in N\backslash S}^{ep} = \frac{a}{2b}$$

We turn now to characterizing the ex-post γ-core; We skip the preliminaries as they were covered in the previuous discussion of the ex-post γ-core without an emissions tax. Instead we concentrate on the balanced game condition with the emissions tax which is given by the following condition:

$$\sum_{S\in C}\delta_S w^{\gamma}(S)$$

$$\le \sum_{S\in C}\delta_S \sum_{i\in S}\left[g\left(e_i^{ep}\right) - te - v\left(\sum_{j\in N\in} e_j^{ep}\right)\right]$$

$$= \sum_{i\in N}\left[g\left(e_i^{ep}\right) - te - v\left(\sum_{j\in N\in} e_j^{ep}\right)\right]\sum_{S\ni i}\delta_S$$

$$= \sum_{i\in N}\left[g\left(e_i^{ep}\right) — te - v\left(\sum_{j\in N\in} e_j^{ep}\right)\right] \ge \sum_{i\in N} g(\widehat{e_\imath})$$

We know that for a zero tax rate the final inequality in the expression is strict. However a positive tax rate will reduce the left-hand side of this inequality and if chosen to be sufficiently large could result in equality. As we are free to choose the tax rate through policy, this condition suggests choosing the emissions tax to balance the game such that

$$\sum_{S\in C}\delta_S w^{\gamma}(S) \le \sum_{i\in N} g(\widehat{e_\imath}) = v(N)$$

This will be the case if one chooses a tax rate such that:

$$t > \frac{\sum_{i\in N}\left[g\left(e_i^{ep}\right) - v\left(\sum_{j\in N} e_j^{ep}\right)\right] - \sum_{i\in N} g(\widehat{e_\imath})}{\sum_{i\in N} e_i^{ep}}$$

In other words taxes per unit of emissions need to be set at least as large as the ratio of aggregate need benefits from deviating from the agreement ex-post less aggregate benefits of a successful agreement per unit of aggregate emissions resulting from ex-post deviation. If the tax were set less than this then the benefits from deviating would outweigh the costs of sticking to the agreement and the agreement would fall apart.

A sufficiently high tax rate therefore results in in non-emptiness of the ex-post γ-core. Consequently, there is a policy measure available to stable international environmental agreements ex-post even in the presence of thresholds and a parameter constellation that suggests countries would have an incentive to deviate from successful agreementst. We have constructed this solution so that the tax is part of the initial agreement, in other words an tax is imposed ex-ante to deal with a problem that might emerge ex-post. This avoids "band-aid" solutions such as only imposing the tax ex-post to deal with deviations once they are observed. As a result the tax will also have an ex-ante impact in that emissions will be reduced due to the tax, this means that negotiated emissions reductions need be less stringent if coupled to other emissions reduction mechanisms. The tax will support emissions reductions efforts rather than hinder them.

It is worth noting however that not every country may be willing to impose private sanctions on its citizens in order to guarantee a successful outcome. The unconstitutionality of the proposed French carbon tax springs to mind. However even if a small number of countries were to impose such sanctions it may be sufficient to reduce ex-post emissions below threshold levels. Consequently, robust and successful IEA's are likely to require a mix of both transfer mechanisms designed to gain agreement to agree to emissions reductions as well as some type of private sanctions, mostly likely, tax instruments, in order to guarantee that successful agreements do not evaporate once they have served their purpose.

7. Conclusion

This paper presents a model of IEAs under the assumption that the global climate is an environmental threshold good. The agreement involves consensus between nations as to the level of abatement of greenhouse gas emissions and how the net benefits are to be shared. The model is used to examine incentives and disincentives for signing agreement due to a critical number of other players consenting to commit. The model is used to represent an IEA as a provision point mechanism and explore implications for the number of signatories needed to successfully reduce emissions below a harmful threshold. The central issue in climate change policy is to bring parties to a consensus regarding the level of harmful ambient greenhouse gas concentrations while the business as usual case continue to build the greenhouse gas levels with the increasing potential for concentrations to reach the unknown critical threshold. Not taking action then predisposes the global community an increasing risk of damage.

We examined whether a high threshold - an emissions level that is unlikely to be reached, could lead to a smaller group of signatories in equilibrium? Then, can a partial agreement equilibrium induce emissions levels that drive the ambient greenhouse gas concentration below the provision point? This is really the key practical question, because if this is not the case, then IEA will be ineffective in eradicating damage.

Our findings allow us to conclude that a higher threshold for pollutants would in equilibrium result in a lower minimum threshold for membership of the coalition in an agreement that reduces emissions below damaging levels. This should be interpreted to mean that it would be hard to gain signatories to agreements if the damage threshold is high. Low thresholds which are easily crossed on the other hand and the success of which is easily achieved are easier to obtain commitment for a binding agreement. The downside of such agreements is that they may well not be robust as individual signatories may chose to defect because the benefits are shared by all nations whether or not they sign the agreement.

Our analysis complements that of Fuentes-Albero and Rubio (2010) who found that heterogeneity between countries has no relevant effects on the scope of environmental cooperation in comparison with the homogeneous case if transfers are not allowed. With transfers, effects depend on the kind of asymmetry. If abatement costs are different, only limited cooperation can be bought through transfers. On the contrary, if the countries differ in terms of environmental damages, the level of cooperation increases with the degree of asymmetry.

We aim to extend our analysis to consider how potential high abatement costs for developed nations can be a factor in inducing greater international cooperation.

References

Barrett, S. (1994) Self-enforcing international environmental agreements, *Oxford Economic Papers* 46: 878-894.

Cararro, C., Marchiori, C. and Oreffice, S. (2004) Endogenous minimum participation in international environmental treaties, Centre for Economic Policy Research, Discussion paper no. 4281.

Cararraro, C. and Siniscalco, D. (1993) Strategies for the international protection of the environment, *Journal of Public Economics* Vol. 52, no. 3., pp. 309-328.

Bergemann, D. and Morris, S. (2004) Robust mechanism design, *Econometrica,* Vol. 73, No.6, pp. 1771-1813.

Bergemann, D. and Morris, S. (2008) Ex-post implementation, *Games and Economic Behavior*, Vol. 63, pp. 527-566.

Bergemann, D., & Morris, S. (2009). Robust Implementation in Direct Mechanisms. *Review of Economic Studies, 76*(4), 1175-1204.

Chander, P. (2007) The gamma-core and coalition formation, *International journal of game theory* 35:539–556.

Chander, P. and Tulkens, H. (1997) The core of an economy with multilateral environmental externalities, *International Journal of Game Theory* 26:379-401.

Chander, P. and Tulkens, H. (2006) Cooperation, stability and self-enforcement in international environmental agreements: A conceptual discussion, CORE DISCUSSION PAPER N° 2006/03.

Fuentes-Albero, C., & Rubio, S. J. (2010). Can international environmental cooperation be bought? *European Journal of Operational Research, 202*(1), 255-264.

Helm,C. (2001) On the existence of a cooperative solution for a coalitional game with externalities, *International Journal of Game Theory*, 30: 141-146.

McQuillin, B. (2008) The extended and generalized Shapley value:simultaneous consideration of coalitional externalities and coalitional structure, *MPRA* paper No. 12409, December 2008.

Osborne, M. and Rubinstein, A. *A course in game theory*, MIT press 1994.

Pavlova, Y. (2008) Multistage coalition formation game of a self-enforcing international environmental agreement, Jyväskylä studies in computing, No. 94, University of Jyväskylä.

R.M.Thrall and W.F. Lucas (1963) N-person games in partition function form, *Naval Research Logistics Quarterly* 10, pp. 281-298.